Performance Analysis of Attributes for Handmade Paper Classification Using Information Theory

Shouji Sakamoto（Ryukoku University）

When considering the attributes of paper, it is possible to include a wide variety of characteristics, ranging from relatively easy-to-measure factors such as paper size and weight to the identification of raw materials, fiber arrangement patterns, watermark patterns, transmission, absorption, and reflection patterns of certain electromagnetic waves. However, many of these attributes pose difficulties in measurement.

In the following, we will initially focus on relatively popular paper attributes (except for fiber arrangement) listed in Table, as a starting point for this type of research. Here, let P represent a set of a large number of papers, and S be a set consisting of all molds (papermaking screens) used to make the elements (papers) in P. WM denotes the set of molds with watermark patterns. $|\ |$ represents the cardinality of the set.

Attributes	Codomain Data type	Unit	Range (roughly)	Classification Capability (Entropy)	measurability	spatial resolution for observation	Descriptions
Height	**R**	cm	10s〜100s	Low	easy	cm	
Width	**R**	cm	10s〜100s	Low	easy	cm	
Weight	**R**	g		Low	easy		dense data
Thickness	**R**	mm	< 0.3	Low	easy	< mm	dense data
Area	**R**	cm2	1000s ~ 10000s	Low-Middle	easy		
Volume	**R**	cm3	10s ~ 100s	Low	easy		dense data
Grammage	**R**	g/m²	10s〜100s	Low	easy		
pH	**N**		c. 3〜9	Low	easy		poor range
Laid line density	**N** or **R**	lines/cm	c. 3〜11	Low	easy	mm	dense data
Chain line width	**R**	cm	c. 1〜10	Low	easy	cm	dense data
Fiber type	**Ch**. (words)		c. 10s (*kozo*, flax, etc.)	Low	easy	μm	poor range
Filler type	**Ch**. (words)		c. 10s (wheat, rice starch, etc.)	Low	easy	μm	poor range
Macromolecule type (FTIR, Raman, Chromatography, etc.)	**Ch**. (words)		c. 10s (cellulose, hemicellulose, amylose, lignin, pectin, etc.)	Low	depends	< μm	poor range
Mineral type (XRF analysis)	**Ch**. (words)		c. 10s (Ca, Cl, K, etc.)	Low	easy	< μm	poor range
Paper color	**Ch**. (words)		c. 10s (white, yellow, etc.)	Low	easy	cm	poor range
Paper color (RGB, $L^*a^*b^*$ etc.)	**R**^3		L^*: 70〜100 a^*: -1〜5 b^*: 7〜20	Low	easy	mm	depends on light source
Number of fibers	**N**		? (depends on paper size)	Middle	hard	μm	
C14/C12 (Carbon dating)	**N/N**		? (depends on quantity of sample)	Middle	easy	< μm	
Laid lines pattern	2D array data (image)		< $\|S\|$	Low	easy (depends)	mm	dense data
Chain lines pattern	2D array data (image)		< $\|S\|$	Middle	easy (depends)	cm	sparse data
Papermaking screen pattern	2D array data (image)		< $\|S\|$	Middle	easy (depends)	mm	sparse data
Watermark (as motif)	**Ch**. (words)		1000s (< $\|WM\|$)	Middle	easy (depends)	cm	thousands words
Watermark (as pattern)	2D array data (image)		$\|WM\|$	Middle	easy (depends)	cm	sparse data
Paper color pattern	2D array data (image)		c. $\|P\|$?	High	easy	< mm	depends on light source
Sheet formation pattern	2D array data (image)		c. $\|P\|$?	High	easy	< mm	depends on light source
Thickness pattern	2D array data (image)		c. $\|P\|$?	High	hard	< mm	
Fiber arrangement	3D array data		$\|P\|$	High	hard	μm	

The laid lines pattern refers to a pattern in the horizontal direction created by bamboo sticks and similar materials, excluding the chain lines pattern. The chain lines pattern is a pattern in the vertical direction created by threads. Papermaking screen pattern refers to both the pattern formed by bamboo strips (or reed etc.) and threads that make up the mesh.

- *Codomain*: **R**: Set of real numbers, **N**: Set of natural numbers, **Ch**.: Set of character strings, 2D array data: Set of 2-dimensional data (e.g., images), 3D array data: Set of 3-dimensional data (e.g., point cloud data).

  ∗Note: Although labeled as 2D array data or 3D array data, the attributes are not 2-dimensional or 3-dimensional data themselves. For example, RGB image data with dimensions of 1920×1080 pixels represents a 6220800 (= 1920×1080×3) dimensional vector data.
- *Range*: The range of attribute values in the specified units (values in the table are approximate).
- *Classification Capability*: The level of detail in classification based on each attribute, considering the presence of a large number of papers.

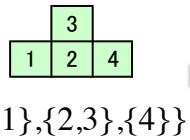
Partitions and ordering relations of set U

Let U={1, 2, … , *i*, … , *n*} be a finite set of *n* papers.
※To simplify the notation, the papers are expressed as natural numbers.

Let *f* be a function that returns the attribute value of *i*∈U.
[There are an infinite number of such functions *f*]

i∈U	*f*(*i*)
1	a
2	b
3	b
4	c

Attribute function *f*



{{1},{2,3},{4}}

Partition of a set U by a function *f*

$$\left(\frac{|\{1\}|}{|U|},\frac{|\{2,3\}|}{|U|},\frac{|\{4\}|}{|U|}\right)$$

$$\left(\frac{1}{4},\frac{2}{4},\frac{1}{4}\right)$$

Probability distribution

$$H=-\sum_{i\in U}p_i\log_2 p_i=\left(-\frac{1}{4}\log_2\frac{1}{4}\right)+\left(-\frac{2}{4}\log_2\frac{2}{4}\right)+\left(-\frac{1}{4}\log_2\frac{1}{4}\right)=1.5$$

Entropy
Information value of *f*

● The attribute function *f* generates a partition of the set U.

●Stirling number of the second kind *S*(*n*, *k*)

The number of ways to partition a set of *n* papers into *k* non-empty classes

$$S(n,k)=\frac{1}{k!}\sum_{i=0}^{k}(-1)^{k-i}{}_kC_i\,i^n$$

●Bell number *B*(*n*)

The number of the possible partitions of a set of *n* papers

$$B(n)=\sum_{k=0}^{n}S(n,k)$$

●Shannon entropy *H*

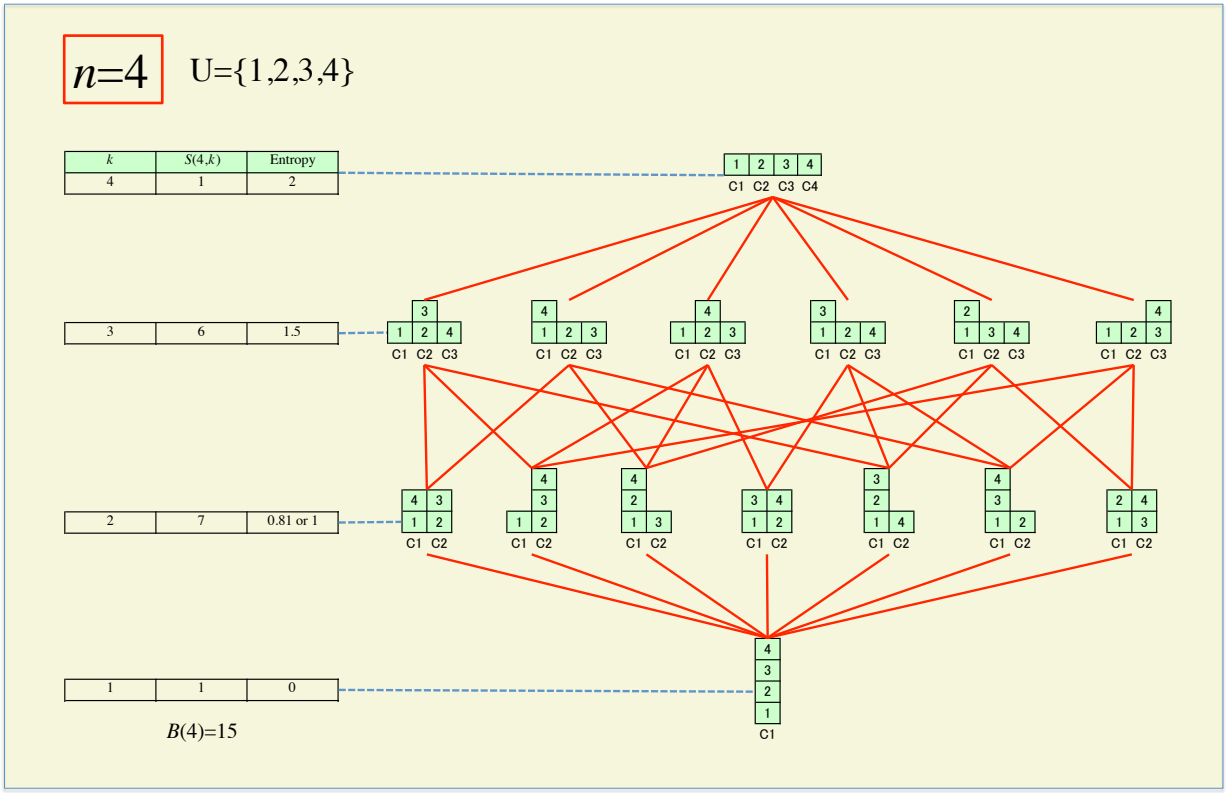
The Information value derived from probability distribution evoked by partitions of set U.
[Measures of complexity, diversity, minimum code length, etc.]

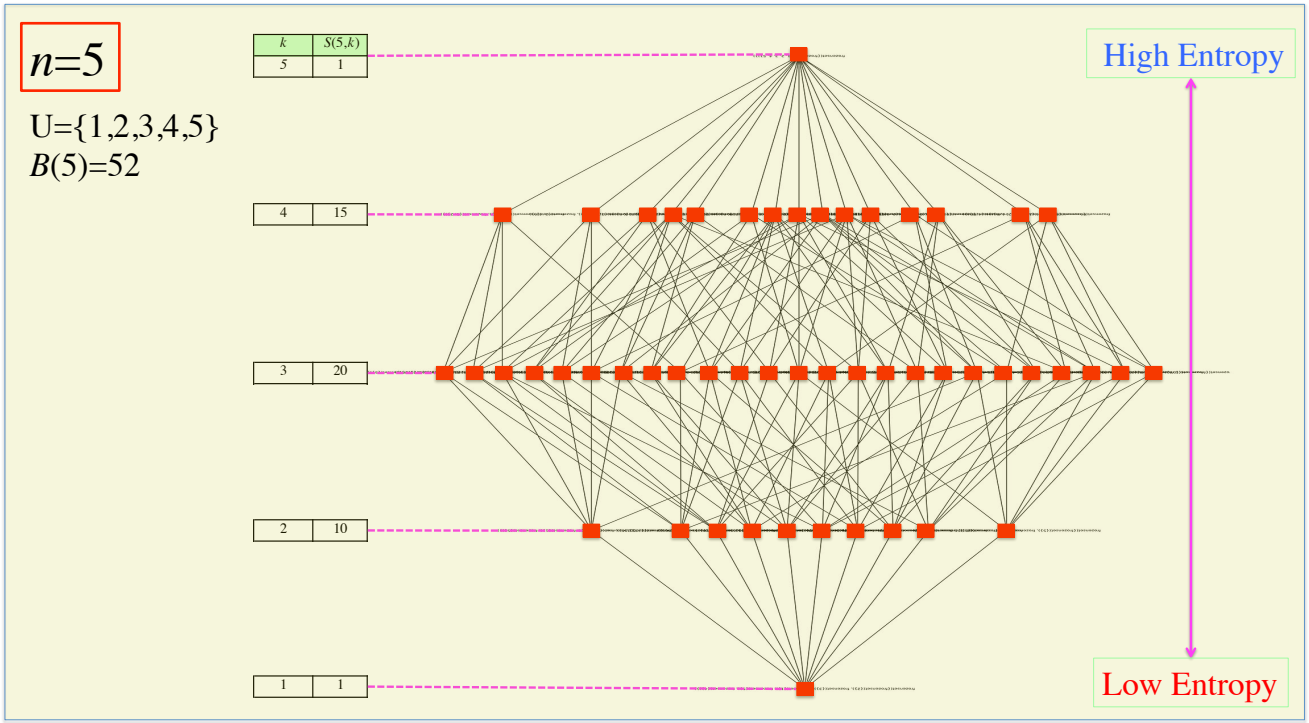
$$H=-\sum_{i=0}^{k}p_i\log_2 p_i$$

●Partition function *P*(*n*)

The number of possible partitions of a positive integer *n*
[The total number of possible entropy values]

n	1	2	3	4	5	6	7	8	9	10	…	20	…	50	…	100	…
P(*n*)	1	2	3	5	7	11	15	22	30	42	…	627	…	204226	…	190569292	…





	Max Entropy	0	1	log3	2	log5	log6	log7	3	log9	log10	…	log$k$	…	log$n$	…
$n = \|U\|$	$B(n)$	$S(n,1)$	$S(n,2)$	$S(n,3)$	$S(n,4)$	$S(n,5)$	$S(n,6)$	$S(n,7)$	$S(n,8)$	$S(n,9)$	$S(n,10)$	…	k	…	n	…
1	1	1														
2	2	1	1													
3	5	1	3	1												
4	15	1	7	6	1											
5	52	1	15	25	10	1										
6	203	1	31	90	65	15	1									
7	877	1	63	301	350	140	21	1								
8	4140	1	127	966	1701	1050	266	28	1							
9	21147	1	255	3025	7770	6951	2646	462	36	1						
10	115975	1	511	9330	34105	42525	22827	5880	750	45	1					
11	678570	1	1023	28501	145750	246730	179487	63987	11880	1155	55	1				
12	4213597	1														
13	27644437	1														
14	190899322	1														
15	1382958545	1														
16	10480142147	1														
17	82864869804	1														
18	682076806159	1														
19	5832742205057	1														
20	51724158235372	1	524287	580606446	45232115901	749206090500	4306078895384	11143554045652	15170932662679	12011282644725	5917584964655	…	$S(20,k)$	…		
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	1.85724E+47	1	562949953421311													
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	4.75854E+115	1	6.33825E+29	8.58963E+46	6.69558E+58	6.57384E+67	9.07387E+74	6.41760E+80	5.05211E+85	7.31910E+89	2.75500E+93	…	$S(100,k)$	…		
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n		1											$S(n,k)$		1	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Paper Classification Performance based on Attributes.

Attributes	Codomain Data type	Unit	Range (roughly)	Classification Capability (Entropy)	measurability	spatial resolution for observation	Descriptions
Height	**R**	cm	10s～100s	Low	easy	cm	
Width	**R**	cm	10s～100s	Low	easy	cm	
Weight	**R**	g		Low	easy		dense data
Thickness	**R**	mm	< 0.3	Low	easy	< mm	dense data
Area	**R**	cm2	1000s ~ 10000s	Low-Middle	easy		
Volume	**R**	cm3	10s ~ 100s	Low	easy		dense data
Grammage	**R**	g/m²	10s～100s	Low	easy		
pH	**N**		c. 3～9	Low	easy		poor range
Laid line density	**N** or **R**	lines/cm	c. 3～11	Low	easy	mm	dense data
Chain line width	**R**	cm	c. 1～10	Low	easy	cm	dense data
Fiber type	**Ch**. (words)		c. 10s (*kozo*, flax, etc.)	Low	easy	µm	poor range
Filler type	**Ch**. (words)		c. 10s (wheat, rice starch, etc.)	Low	easy	µm	poor range
Macromolecule type (FTIR, Raman, Chromatography, etc.)	**Ch**. (words)		c. 10s (cellulose, hemicellulose, amylose, lignin, pectin, etc.)	Low	depends	< µm	poor range
Mineral type (XRF analysis)	**Ch**. (words)		c. 10s (Ca, Cl, K, etc.)	Low	easy	< µm	poor range
Paper color	**Ch**. (words)		c. 10s (white, yellow, etc.)	Low	easy	cm	poor range
Paper color (RGB, $L^*a^*b^*$ etc.)	**R**^3		L^*: 70～100 a^*: -1～5 b^*: 7～20	Low	easy	mm	depends on light source
Number of fibers	**N**		? (depends on paper size)	Middle	hard	µm	
C14/C12 (Carbon dating)	**N/N**		? (depends on quantity of sample)	Middle	easy	< µm	
Laid lines pattern	2D array data (image)		< | *S* |	Low	easy (depends)	mm	dense data
Chain lines pattern	2D array data (image)		< | *S* |	Middle	easy (depends)	cm	sparse data
Papermaking screen pattern	2D array data (image)		< | *S* |	Middle	easy (depends)	mm	sparse data
Watermark (as motif)	**Ch**. (words)		1000s (< | *WM* |)	Middle	easy (depends)	cm	thousands words
Watermark (as pattern)	2D array data (image)		| *WM* |	Middle	easy (depends)	cm	sparse data
Paper color pattern	2D array data (image)		c. | *P* | ?	High	easy	< mm	depends on light source
Sheet formation pattern	2D array data (image)		c. | *P* | ?	High	easy	< mm	depends on light source
Thickness pattern	2D array data (image)		c. | *P* | ?	High	hard	< mm	
Fiber arrangement	3D array data		| *P* |	High	hard	µm	

f_L

f_M

f_H

U: finite set of papers

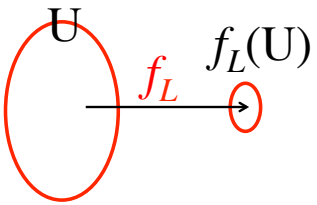
Let f be a function that returns the attribute value of $p \in U$.

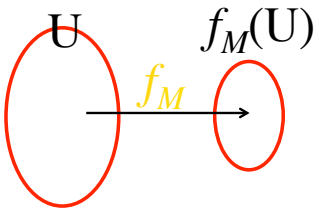
Low entropy attribute

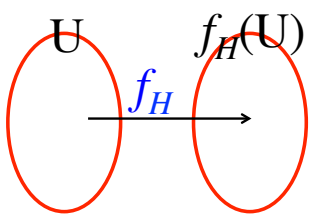
Attributes that are expressed in low-dimensional values or vocabulary, and that appear in natural language descriptions, are easy for humans to understand but have low partitioning performance for U.

High entropy attribute

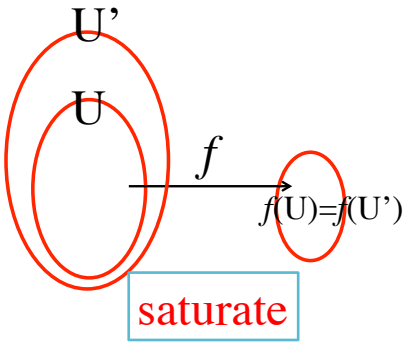
Attributes expressed as high-dimensional data, which are difficult to describe in natural language due to limited vocabulary, have high partitioning performance for U.

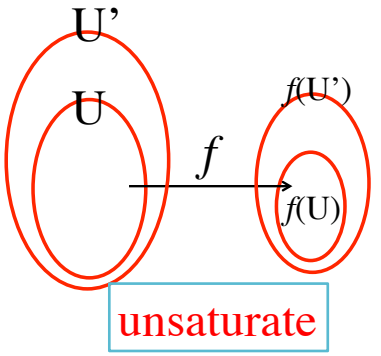




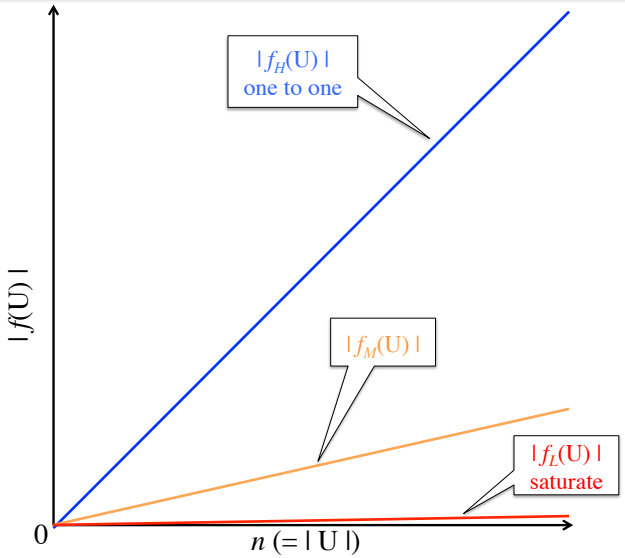


$$\left|f_L(U)\right| < \left|f_M(U)\right| < \left|f_H(U)\right|$$

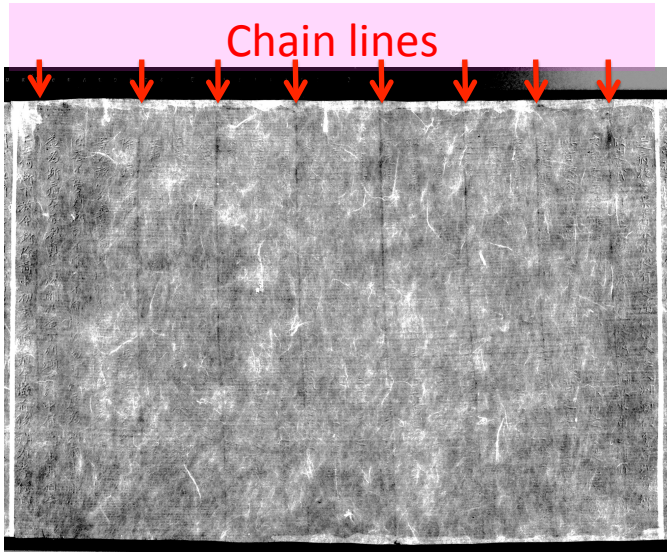




If we consider increasing the size of the set U



The change in the number of partitions of U according to each attribute function as U varies.



Papermaking screen pattern

High entropy attributes

・2D and 3D array data (high dimensional data)
[Looking closer gives more information, Redundant description]

● The attribute “fiber arrangement” shows the highest classification capability as it allows for the fine classification of each paper. This is because papers with exactly the same arrangement of fibers are almost non-existent. However, this attribute corresponds to high-dimensional, large-scale data, such as high-resolution images.

・Low dimensional data [Comprehensive overview]

※ While assigning a unique identifier to each sheet of paper or recording its exact spatial coordinates can allow for the classification of all sheets, such metadata does not capture the internal information inherent to the paper itself.

Middle entropy attributes

・Time-varying attribute
Replacement due to lifespan (e.g. papermaking screen, ^{14}C, (papermaker))
[More information will be available over time]

● The high classification capability of the screen pattern, including chain line pattern and/or watermark, attribute
※ The attribute "*chain line pattern*" is a pattern created by threads that connect the arranged reed or bamboo strips. While each thread appears to be similar to bamboo strips in terms of thickness, the spatial freedom in thread arrangement significantly differs from the laid lines pattern, resulting in a high diversity of the chain line pattern and a high classification capability. Even when attempting to make the chain line width uniform in a mesh configuration, it is not easy to maintain equal intervals between all threads during manual weaving by hand, resulting in differences in thread spacing (about a few millimeters). Such spacing differences are easily noticeable to the naked eye.
※ Although watermark analysis in Western paper involves an extensive vocabulary (e.g., "foolscap," "*tre lune*"), these terms originally stem from unrelated fields rather than intrinsic paper-related terminology.

Low entropy attributes (saturate)

Why Saturate?
・Limited material (vocabulary)
・Low dimensional data with lower and higher bounds by human body size
・Measurement accuracy limits
[Information loss]

● Attributes with the codomain of **Ch**. (words) (e.g. fiber type such as *kozo*, flax, cotton, bamboo, filler type such as wheat, millet, rice) also have low classification capability. This is due to the limited range of attribute values.
(※For example, individual flax fibers are not regarded as distinct entities, but rather are treated as identical or equivalent to one another.)
● Attributes with the codomain of **R**, **N** (height, width, weight, paper thickness, area) have low classification capability. This is because there are many papers with similar attribute values, usually.

Summary

(1) Even when analyzing the simple physical properties or conducting scientific material analyses of paper, most of the resulting attribute values suffer from significant loss of original information and tend to saturate at a fixed number of measurable properties.

Most scientific material identification methods—such as XRF, FTIR, Raman spectroscopy, and GC-MS—can detect only a limited range of materials from paper, demonstrating a saturation effect in their analytical capacity.

(2) Furthermore, the number of attributes typically investigated remains limited, often around 10, 20, or 30 at most.

As a result of points (1) and (2), the number of possible classifications becomes saturated. As the number of paper samples to be classified increases, the number of possible classifications theoretically grows exponentially (according to the Bell number), but current methods cannot accommodate such growth.

(3) The only large-scale successful analytical method for paper to date has been watermark analysis, which has a history of over 100 years. In contrast, for papers without watermarks—such as those from Arab or East Asian traditions—the chain line pattern has a classification potential comparable to watermarks. However, this approach has received little attention and remains largely unexplored.

The uniqueness of papermaking screens—since virtually no two are identical—and their eventual replacement due to deterioration over time mean that screen-mate papers inherently retain information indicating that they were produced in the same location within a limited timeframe (i.e., during the operational period of a given screen). However, it is not necessarily possible to determine the absolute time or location.

(4) An attribute function that assigns a unique name or identifier to each sheet of paper can theoretically describe all possible partitions of the set U of papers (i.e., it has universality and forms a basis). However, such a function contains no information about the intrinsic characteristics of the paper itself.

In contrast, an attribute function that assigns high-dimensional vector values to each paper—based on observational data (e.g., high-resolution images, or visual information derived from the retina in human perception)—possesses both high descriptive redundancy and a rich informational content. [Humans tend to think in terms of significantly reduced representations of such complex information, often replacing it with abstract concepts or low-dimensional numerical values that entail substantial information loss.]