THE XPRIORI REPORT

An Actual Case on Unlocking Value in Information – A Large Petroleum Company Uses of Data Similarity Technology in Large Scale Classification Projects
Part 4

Concepts of Similarity or Sameness are Critical to Human Understanding; A Technology that Automates Clustering/Classification by Similarity is a BIG DEAL!; A Technology that can continually automate classification without much by way of human intervention is an even BIGGER DEAL!

I. Introduction

As we indicated in Parts 1-3, in recent months, a large Petroleum Company has launched a series of projects using new automated document clustering technologies to evaluate and organize these largely digital information assets with a view to creation of value and competitive edge. The Company is undergoing a revamp of its enterprise content management strategy and philosophy at the functional and business asset level. As is the case with many organizations, the primary objectives of this exercise include enterprise-wide cost containment, risk reduction and the extraction of value from content while enabling more effective use and management of the asset.

In Part 1, we discussed the goals and values for the project; the need for new technologies to supplant the slow pace of manual review and the small but expert team that was capable of using the new technologies to meet those goals, values and objectives. In Part 2, we addressed the team's ability to meet the overarching goals for the system of *Accuracy, Findability, Consistency and Governance* at the various stages of creation, use and storage of information at the Company. We also offered a number of keys to understanding the process and deployment of our technology. In Part 3, we addressed the two phase work flow deployed at the Company and provided some details of the step by step process that was followed.

In this Part 4, we now discuss how "similarity" has been and is the basis for assessment or categorizing all sorts of phenomena and things for most of history, and how it is the basis for classification today even in the new digital settings. We also consider again the main value propositions for our approach in this new era.

II. Similarity As The Basis For Classification

Assessments of similarity seem to be important for a variety of cognitive acts, ranging from problem solving to categorization to memory retrieval. William James (1890/1950) was correct in stating that "this sense of sameness is the very keel and backbone of our thinking". This statement could appear

in any number of articles in scholarly journals dealing with vector and other algorithmic analysis designed to assess similarity of documents, events, or anything expressed within spatial bounds.

When one perceives similarity of various things, expressions or phenomena, one tends to use it as suggesting a context for what makes them "similar". Context adds meaning to information or even idle expressions. In fact, context is necessary for us to have information. Think about it. I can say the word "Tim" and it means little or nothing. However, if I say "My name is Tim," the word Tim now has a useful context from which the word can be understood, used or consumed. Without the context, the term would rest on a stack of total miscellany to be ignored until a suitable context is found or provided.

In the case of document classification, eDiscovery, information governance etc., we tend to be dealing with aligning content with events or business activities that, at least in a larger context, have a degree of commonality at some level of understanding or presentation. There is less of a chance of finding the total chaos of, say, 1,000,000 unrelated or totally dissimilar expressions of content, where the content has been created in the course of a particular activity, such as in this case, the finding, drilling for and producing fossil fuels and/or related support activities. In this context, finding 1,000,000 totally dissimilar documents would be unexpected.

One of my colleagues argues that all electronic data that is created anywhere in the world has a threshold of similarity >0% to every other known data type that exists (when one considers metadata in addition to the substantive content). I guess that I am more willing to accept that electronic data created in the context of a business or functional activity will more likely have some threshold of similarity >0%. I can't prove it directly as I write this monograph, but I am willing to advance it for more than argument's sake here. This is particularly the case where we are looking for and grouping items based upon "similarity" and then submitting the clusters to users (Document Controllers) for confirmation that they have some useful degree of similarity. Users decide to confirm or not, the proposed associations based upon your knowledge, education, experience and likely some consensus arising from group discussion. The "machine" has only enabled or processed the associations considering all of the content presented in the data set.

III. Automating Clustering By Similarity Saves Time, Money and Speeds Process, and Enables One That is More Accurate Than Purely Human Manual Review

Why is all of this important? Essentially, there are four primary reasons : (a) the machine enabled clustering takes into consideration every document in the data set; (b) one can reasonably classify and/or cull documents by reviewing less than all documents in a cluster, without having read each document; (c) the algorithmically based process takes into consideration non-text based information; and (d) the underlying code that created the clusters is preserved and is automatically applied to information newly introduced to a particular dataset. Let's take a look at each of these primary reasons.

(a) Consider a process that looks at every document. Many of us possess experience based intuition that will produce thoughtful and productive Boolean searches. If one knows the territory, he can do a pretty good job with those searches… but we are deploying to experientially developed guesses that are not directly derived from the content we desire to assess. Useful? … Well, yes. But, are conclusions drawn from an assessment of all of the content presented? … Well, no. The same is true for predictive coding. We can spot exemplars with some alacrity; however, there is no guarantee that those exemplars will lead to an assessment of every document in the set. Our clustering process does.

(b) Consider information a cluster at a time; not a document at a time. It is easy to think of culling of clusters of third party information – say, groups of periodicals. It is just as easy to recognize after reading several documents in a cluster the substantive basis for their being clustered together and stored based upon a header in a taxonomy. At the Petroleum Company, both activities occurred. There was an existing taxonomy that was available. However, in the process of cluster review, it was substantially updated – more than doubling the number of classification points in the existing taxonomy. The results of the largely automated clustering processes, both textual and visual in nature, caused the team to suggest significant changes to the taxonomy and its hierarchical presentation. The completeness of the existing taxonomy was assessed while new clusters simultaneously suggested a basis for enabling content/data driven changes to it.

Figures 1 and 2 below illustrate the before and after result of data analysis. Figure 1 is the "pre-Copernican" view of the data taxonomy as we knew it. Figure 2 is the result of our using the new clusters of data to add additional categories to the taxonomy, providing a greater degree of granularity in the organization of the content.

Doc_Type	System_Abbreviation	Content_Type	Content_Subtype	Original Taxonomy Designation
Agreements	AGR0001	Contract Agreements		Original Taxonomy Designation
Agreements	AGR0002	Consulting Agreements		Original Taxonomy Designation
Agreements	AGR0003	Non Disclosure/ Confidentiality Agreements		Original Taxonomy Designation
Agreements	AGR0004	Advisory Agreements		Original Taxonomy Designation
Agreements	AGR0005	Master Services Agreements		Original Taxonomy Designation
Agreements	AGR0006	Service Agreements		Original Taxonomy Designation
Agreements	AGR0007	Purchase and Sale Agreements		Original Taxonomy Designation
Agreements	AGR0008	Transition Services Agreements		Original Taxonomy Designation

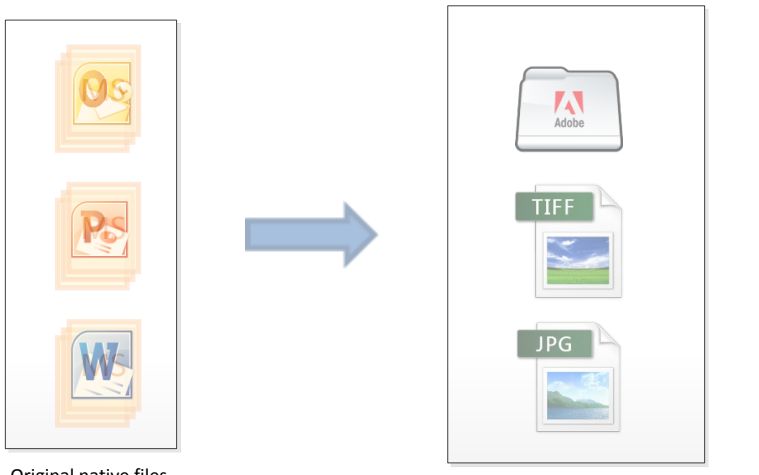
Figure 1 – Illustration of the taxonomy before clustering – Content Sub-type is blank for all records.

Agreements	AGR0001	Contract Agreements		Original Taxonomy Designation
Agreements	AGR0002	Consulting Agreements		Original Taxonomy Designation
Agreements	AGR0003	Non Disclosure/ Confidentiality Agreements		Original Taxonomy Designation
Presentations	AGR0003a	Confidentail Information Memorandum (CIM)		Post Analysis Taxonomy Designation
Agreements	AGR0004	Advisory Agreements		Original Taxonomy Designation
Agreements	AGR0005	Master Services Agreements		Original Taxonomy Designation
Agreements	AGR0006	Service Agreements		Original Taxonomy Designation
Agreements	AGR0006a	Services Agreement	Software	Post Analysis Taxonomy Designation
Agreements	AGR0006b	Service Agreement	Gas	Post Analysis Taxonomy Designation
Agreements	AGR0006c	Services Agreement	Delegated Reporting	Post Analysis Taxonomy Designation
Agreements	AGR0006d	Services Agreement	Amendment	Post Analysis Taxonomy Designation
Agreements	AGR0007	Purchase and Sale Agreements		Original Taxonomy Designation
Agreements	AGR0007a	Purchase and Sale Agreement	Natural Gas	Post Analysis Taxonomy Designation
Agreements	AGR0007b	Purchase and Sale Agreement	Master Power	Post Analysis Taxonomy Designation
Agreements	AGR0007c	Purchase and Sale Agreement	Collateral Annex	Post Analysis Taxonomy Designation
Agreements	AGR0007d	Purchase and Sale Agreement	Crude or Oil	Post Analysis Taxonomy Designation
Agreements	AGR0007e	Purchase and Sale Agreement	Gas Annex	Post Analysis Taxonomy Designation
Agreements	AGR0007f	Purchase and Sale Agreement	Renewable Energy Certificate	Post Analysis Taxonomy Designation
Agreements	AGR0007g	Purchase and Sale Agreement	Consent Agreement	Post Analysis Taxonomy Designation
Agreements	AGR0007h	Purchase and Sale Agreement	Sharing Agreement	Post Analysis Taxonomy Designation
Agreements	AGR0007i	Purchase and Sale Agreement	Electricity	Post Analysis Taxonomy Designation
Agreements	AGR0008	Transition Services Agreements		Original Taxonomy Designation

Figure 2 – Post clustering and analysis – Content Sub-type is populated for new records.

(c) **Deal with both textual and visual similarity to enable consideration of all content.** Many documents are presented in image format and contain non-textual symbols or tokens such as logos, illustrative diagrams, pictures, etc. By converting all files in a document set to image format and then comparing them to selected percentages of similarity, Document Controllers are able to find those that are duplicates to text files and also can now include the non-textual material as part of the clustering process. The process and results are illustrated in the diagrams below. Every day, millions of

substantively identical files get converted into different file types. Even though the PDF or TIFF or JPEG contains the same information as the actual email or PowerPoint or Word document from which it was derived, they have different hash values. This results in the "system" seeing substantively identical files as being duplicates. Figure 3 below illustrates this point.



Original native files

Converted native files (substantively identical to orginals).

Figure 3

Figure 4 below illustrates how visual clustering ignores the file format and focuses on the substance of documents in order to group them based on their similarity with one another. It also illustrates varying results in applying two different percentages of similarity. This visual approach enables organizations to identify substantive duplicates and for the first time, identify potential versions of the same documents that are scattered throughout email and document storage servers. This ability has proved absolutely critical in the engineering, architectural and energy field where it is critical to know that the versions of documents that are being relied upon are the most recent. The results at the Petroleum Company were quite compelling.

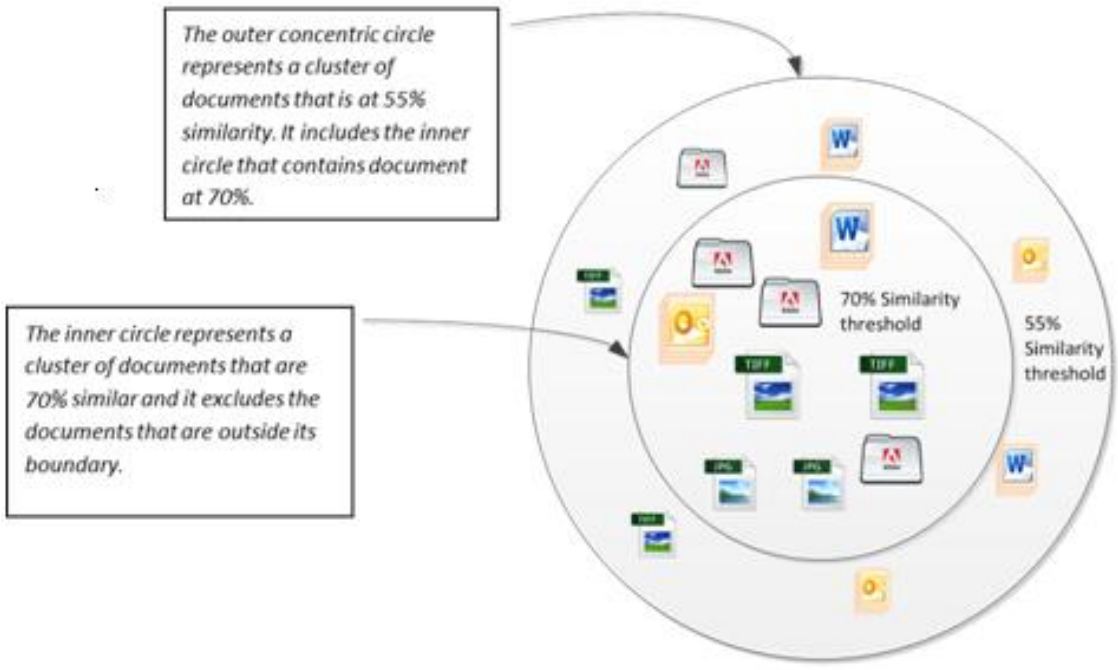


Figure 4

(d) Code generated to create the clusters is preserved and is subsequently applied to newly introduced information. This is the key to implementing an automated self-classification system for new information as it is received at the enterprise or to a program.

New clusters might suggest changes to the taxonomy; however, there is no need to recode for documents that fit to previously created clusters. Projects evolve over time with new information being introduced from time to time for extended periods in the future. Post-closing M & A integration is made far easier. Cluster documents from the acquired company and associate those clusters with those of the acquiring company. The same is true for eDiscovery activities from litigation hold to production. It goes on over time and now you can look at everything that you think might even have a remote chance of containing relevant information.

The implications of this process are significant. Once the classification work previously done throughout the prior Phases is codified, the resulting set of rules can operate independent of human interaction on new information when added. Outliers can still be identified and set aside for further action. As illustrated in Figure 5 below, various departments or other organizational entities in the enterprise can have their file shares monitored by a classification engine which is specifically tuned to their specific classification requirements.

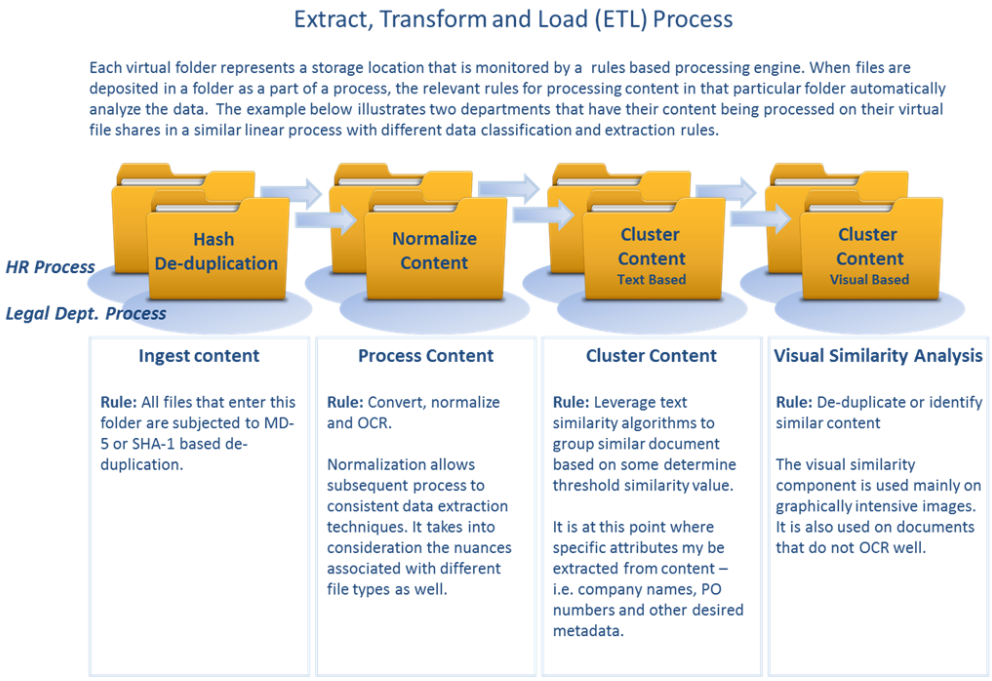


Figure 5

Once classified, the documents can be made available through a common set of access tools, search methodologies and the like to assure that stakeholders can access the information that they require and when they need it. On completion of a classification project in a business unit at the Petroleum Company, all of the foregoing was put in place.

IV. What are the value propositions?

Value Proposition 1 - Data content informs our knowledge of our environment

As illustrated in the foregoing, clustering algorithms now help us understand our data categories in ways heretofore unachievable when working with big data. Clusters can be created without human definition for review and culling or accepting. Much manual effort is avoided and all content is considered. We can now view an organization's enterprise content from a perspective that is completely different from conventional approaches. We no longer have to come at their data from pre-defined point that is, by its nature, grounded in presumptions about the data. Having data describe itself to the user allows the user to see data in its complete context. Data blind spots for which there is no or insufficient classification are revealed in terms of relevance to other known data objects or documents within the corpus of content examined.

Value Proposition 2 – Taming big data

What is big data? Big data is characterized by an acronym describing 3 key variables which has been coined as **V^3**

- **Data Volume** – large data volumes. This is a relative term that changes based on innovations in storage device areal density (the number of bits that can be stored in a given area on a device).
- **Data Variety** – the kinds of data, i.e. unstructured, structured, semi-structured and newer polymorphic content formats.
- **Data Velocity** – the rate at which content is created.

If one accepts the foregoing definition of data, then taming big data is effectively the ability to scale and extrapolate Value Proposition 1 to increasingly large and disparate data volumes. The ability to identify substantively similar and duplicative documents gives organizations the ability to select "the" business records that should be kept in a document archive. The impact on storage budgets can be significant. The amount of storage needed for an organization can now be projected with pinpoint accuracy. The key metrics that allow us to do this can be generated from metadata; storage growth year over year and document duplicates based on clustering.

The impact on understanding and deploying to the right information on a timely basis is even more valuable.

Value Proposition 3 – Sustainable, objective and automated data classification based on iterative clustering and informed and automated modification to the storage taxonomy

The greater the corpus of information that is clustered, the more we know about an organization's document and content types. The process is automated and provides objective review of all content. Each ratified (classified) grouping of document types within an organization becomes the classification exemplar for new documents that enter that particular managed storage environment. For organizations that grow inorganically (by acquisition or merger), the larger the data volumes under management, it is more statistically probable that there will be significant data clusters that will be related to one another irrespective of traditional data horizon impediments such has language or character sets.

Organizations seeking to implement automated and sustainable data classification solutions such as those described and proposed in this document can embark on these types of initiatives with "out of the gate" ROI that positively impacts bottom line. This value proposition has already impacted the following functional business units at the Petroleum Company: (1) Profit centers, (2) Legal departments, (3) Information security groups, (4) Field engineering, (5) HR, (6) and Marketing.

V. Conclusion

Similarity is at the core of how we analyze and categorize information. Our automated clustering is based upon similarity and will cluster documents in highly useful ways. This results in more complete classification of documents and understanding of their importance to the project and the enterprise.