ZyLAB

eDiscovery & Information Management

Text Analysis: The Next Step In Search

ZyLAB White Paper

Johannes C. Scholtes, Ph.D.
Chief Strategy Officer, ZyLAB

Contents

Summary

Text and content analysis differs from traditional search in that, whereas search requires a user to know what he or she is looking for, text analysis attempts to discover information in a pattern that is not known before-hand. One of the most compelling differences with regular (web) search is that typical search engines are optimized to find only the *most* relevant documents; they are not optimized to find *all* relevant documents. The majority of commonly-used search tools are built to retrieve only the most popular hits—which simply doesn't meet the demands of exploratory legal search.

This whitepaper will lead the reader beyond the Google standard, explore the limitations and possibilities of text analysis technology and show how text analysis becomes an essential tool to help process and analyze to-day's enormous amounts of enterprise information in a timely fashion.

Finding Without Knowing Exactly
What to Look For

In general, text analysis refers to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text analysis differs from traditional search in that, whereas search requires a user to know what he or she is looking for, text analysis attempts to discover information in a pattern that is not known beforehand (through the use of advanced techniques such as pattern recognition, natural language processing, machine learning and so on). By focusing on patterns and characteristics, text analysis can produce better search results and deeper data analysis, thereby providing quick retrieval of information that otherwise would remain hidden.

Text analysis is particularly interesting in areas where users must discover new information, such as, in criminal investigations, legal discovery and when performing due diligence investigations. Such investigations require 100% recall; i.e., users can not afford to miss any relevant information. In contrast, a user who uses a standard search engine to search the internet for background information simply requires any information as long as it is reliable. During due diligence, a lawyer certainly wants to find all possible liabilities and is not interested in finding only the obvious ones.

Beyond the Google Standard

With web search engines like Google, most companies and organizations place a premium on being found as close to the top of search list as possible. Experienced users have become quite savvy in utilizing search engine optimization techniques to enhance high rankings.

Now, an entire generation of tech-savvy computer users exist whose expectations and perceptions of full-text search functionality and performance are almost completely influenced by the "Google effect." In most instances, this type of approach works fine if users only need to find the most appropriate website for answering general questions. Users type in full-text keywords and expect to see the most relevant document or website appear at the top of a result list. Page-link and similar popularity-based algorithms work very well in this context.

However, a lot of information that may be vital for them to know may not come to light using only these basic search techniques. If, for example, a user's search is related to fraud and security investigations, (business) intelligence, or legal or patent issues, other searching techniques are needed that support different sets of issues and requirements, such as the following:

- **Focusing on optimized relevance**: the first requirement of broader search applications is that not only does the best document need to be found, but all potentially relevant documents need to be located and sorted in a logical order, based on the investigator's strategic needs.

- **Handling massive data collections**: another issue impacting effective strategic searching is how to conduct extensive searches among extremely large data collections. For example, if email collections need to be investigated, these repositories are no longer gigabytes in size; rather, they can be a terabyte or more. When handling this volume of data, plain full-text search simply cannot effectively support finding, analyzing, reviewing and organizing all potentially relevant documents.

- **Finding information based on words not located in the document**. In this context, consider investigators who may have some piece of information concerning an investigation but don't necessary know other details they may be looking for. Who is associated with a suspect? What organizations are involved? What aliases are associated with bank accounts, addresses, phone records or financial transactions? Traditional precision-focused, full-text approaches are not going to help users find hidden or obscure information in these contexts.

- **Defining relevancy**: when defining a search's relevance, all factors that could be in play during a specific search instance must be accounted for (in the context of overall goals). Using the investigative example again, consider possible involved parties and what "relevance" would mean to their actual search:

  * Investigators want to comb documents to find key facts or associations (the "smoking gun");

  * Lawyers need to find privileged or responsive documents;

  * Patent lawyers need to search for related patents or prior art;

  * Business intelligence professionals want to find trends and analyses; and

  * Historians need to find and analyze precedents and peer-reviewed data.

All of these instances require not only sophisticated search capabilities but also different context-specific functionalities for sorting, organizing, categorizing, classifying, grouping and otherwise structuring data based on additional meta-information, including document key fields, document properties and other context-specific meta-information. Utilizing this additional information will require a whole spectrum of additional search techniques, such as clustering, visualization, advanced (semantic) relevance ranking, automatic document grouping and categorization.

Challenges Facing Text Analysis

Due to the global reach of many investigations, a lot of interest also exists with text analysis in multi-language collections. Multi-language text analysis is much more complex than it appears because, in addition to differences in character sets and words, text analysis makes intensive use of statistics as well as the linguistic properties (such as conjugation, grammar, tenses or meanings) of a language. A number of multi-language issues will be addressed later in this article.

But perhaps the biggest challenge with text analysis is that increasing recall can compromise precision, meaning that users end up having to browse large collections of documents to verify their relevance. Standard approaches to countering decreasing precision rely on language-based technology, but when text collections are not in one language, are not domain-specific and/or contain documents of variable sizes and types, these approaches often fail or are too sophisticated for users to comprehend what processes are actually taking place, thereby diminishing their control.

Furthermore, according to Moore's Law, computer processor and storage capacities double every 18 months, which, in the modern context, also means that the amount of information stored will double during this timeframe as well. The continual, exponential growth of information means most people and organizations are always battling with the specter of information overload. Although effective and thorough information retrieval is a real challenge, the development of new computing techniques to help control this mountain of information is advancing quickly as well. Text analysis is at the forefront of these new techniques, but it needs to be used correctly and understood according to the particular context in which it's applied. For example, in an international environment, a suitable text analysis solution may consist of a combination of standard relevance-ranking with adaptive filtering and interactive visualization, which is based on utilizing features (i.e. metadata elements) that have been extracted earlier.

Control of Unstructured Information

More than 90% of all information is unstructured, and the absolute amount of stored unstructured information increases daily. Searching within this information, or performing analysis using database or data mining techniques, is not possible, as these techniques work only on structured information. The situation is further complicated by the diversity of stored information: scanned documents, email and multimedia files (speech, video and photos).

Text analysis neutralizes these concerns through the use of various mathematical, statistical, linguistic and pattern-recognition techniques that allow automatic analysis of unstructured information as well as the

extraction of high quality and relevant data. ("High quality" here refers to the combination of relevance [i.e. finding a needle in a haystack] and the acquiring of new and interesting insights.) With text analysis, instead of searching for words, we can search for linguistic word patterns which enable a much higher level of search.

Different Levels of Semantic Information Extraction

Several options exist for extraction and text analysis within its products. These options vary from simple extraction methods such as file and document property extraction to more advanced text analysis options:

- **File system extraction**: extraction of file properties such as file name, file size, modified date, creation date, attributes, mime type, etc.

- **Document property extraction**: extraction of specific document properties depending on the document format such as Title, Author, Publisher, Version, etc.

- **Email property extraction**: extraction of common email properties such as Sender, Recipient, Sent Date, Subject, Conversation topic and other properties such as Internet Headers, Original Sender, etc.

- **Microsoft SharePoint property extraction**: extractions of all Microsoft SharePoint document properties as these are stored in SharePoint with the document including security settings.

- **Hash calculation**: calculation of hash values for identification purposes, supporting several hash types such as MD5 and SHA1.

- **Duplicate detection**: calculating hash values based on the content for email messages or binaries for other file types to find and detect duplicate documents.

- **Language detection**: detection of document language, support for over 400 languages.

- **Concept extraction**: extraction of predefined (full-text) queries that identify document and meta information content with specific combinations of keywords or (fuzzy and wildcard) word patterns in.

- **Entity extraction**: extraction of basic entities that can be found in a text such as: people, companies, locations, products, countries, and cities.

- **Fact extraction**: these are relationships between entities, for example, a contractual relationship between a company and a person.

- **Attributes extraction**: extraction of the properties of the found entities, such as function title, a person's age and social security number, addresses of locations, quantity of products, car registration numbers, and the type of organisation.

- **Events extraction**: these are interesting events or activities that involve entities, such as: "one person speaks to another person", "a person travels to a location", and "a company transfers money to another company".

- **Sentiment detection**: finding documents that express a sentiment and determine the polarization and importance of the sentiment expressed.

- **Extended natural language processing**: Part-of-Speech (POS)tagging for pronoun, co-reference and anaphora resolution, semantic normalization, grouping, entity boundary and co-occurrence resolution.

Steven Paul Jobs (born February 24, 1955) is an American business magnate and inventor. He is well known for being the co-founder and chief executive officer of Apple. Jobs also previously served as chief executive of Pixar Animation Studios; he became a member of the board of The Walt Disney Company in 2006, following the acquisition of Pixar by Disney.

Steven Paul Jobs; Jobs	Person
February 24, 1955; 2006	Date
co-founder; chief executive officer; chief executive; member of the board	Title
Apple; Pixar Animation Studios; Pixar; The Walt Disney Company; Disney	Organization_Commercial
American business magnate; inventor; acquisition	Noun_Group or Concept
Jobs also previously served as chief executive of Pixar…	Executive Job Change
Acquisition of Pixar by Disney	Merger and Acquisition

Figure1: Entities and relations between entities in the form of facts and events are extracted

The example below also automatically recognizes sentiments in emails and other documents. This can be very relevant to find a "smoking gun" and "hidden agendas" faster.
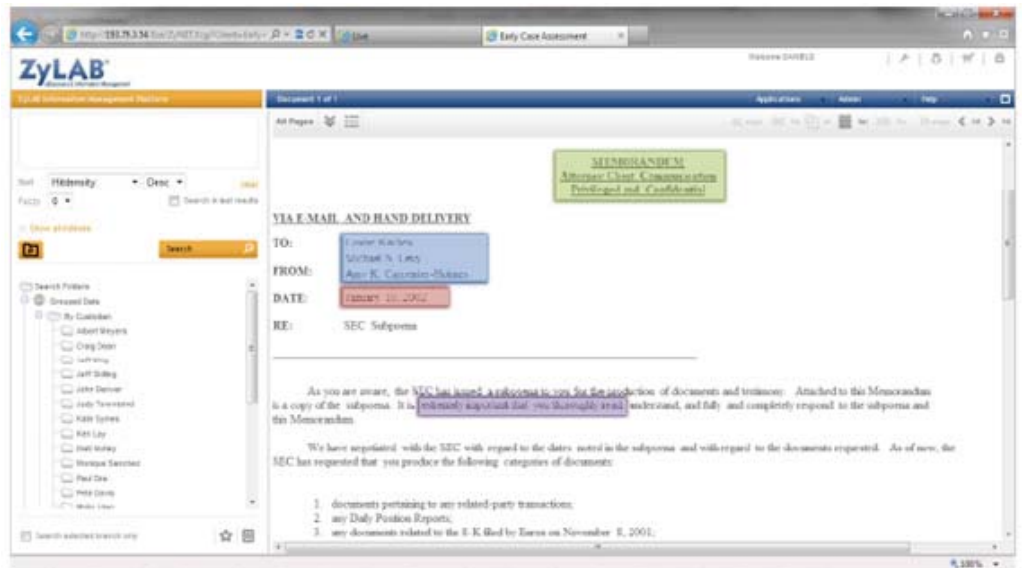
Persons	Amy Carpenter-Holmes, Amy K. Carpenter-Holmes, Lisa Plegge, Louise Kitchen, Michael N. Levy
Privileged	Attorney Client
Sentiment	extremely important that you thoroughly read
Date	January 10, 2002
City	New York, Washington, Washington, DC

Figure 2: Automatic recognition of sentiments

Other examples of the application of text analytics and sentiment mining can be found in the examples below which are related to the FCPA and the UK Bribery Act: in the future it will be more and more important to find potential violations of these acts to prevent expensive investigations, consequential serial litigation and loss of reputation and business!

Figure 3: Searching for risky data combinations

Figure 4: Searching for potential stupidities

Co-reference and Anaphora Resolution

One of the biggest problems in the discovery and identification of events is the resolving of the so called *anaphora* and *co-references*. This is the linguistic problem to associate pairs of linguistic expressions that refer to the same entities in the real world.

Consider the following text:

```
"A man walks to the station and try to
catch the train. His name is John Doe.
Later he meets his colleague, who has just
bought a card for the same train. They
work together at the Rail Company as tech-
nical employees and they are going to a
meeting with colleagues in Washington DC."
```

The text contains are various references and co-references. Various anaphora and co-references will have to be disambiguated before it is possible to fully understand and extract the more complex patterns of events. The following list shows some examples of these (mutual) references:

- **Pronominal Anaphora**: he, she, we, oneself, etc.
- **Proper Name Co-reference**: For example, multiple references to the same name.
- **Apposition**: the additional information given to an entity, such as "Jan Jansen, the father of Piet Jansen".
- **Predicate Nominative**: the additional description given to an entity, for example "Jan Jansen, who is the chairman of the football club".
- **Identical Sets**: A number of reference sets referring to equivalent entities, such as "Ajax", "the best team", and the "group of players" which all refer to the same group of people.

With advanced computational linguistics, one can resolve co-references, pronouns and other anaphora and easily find 2 to 4 times more relevant patterns which dramatically improves the quality of these types of analyses.
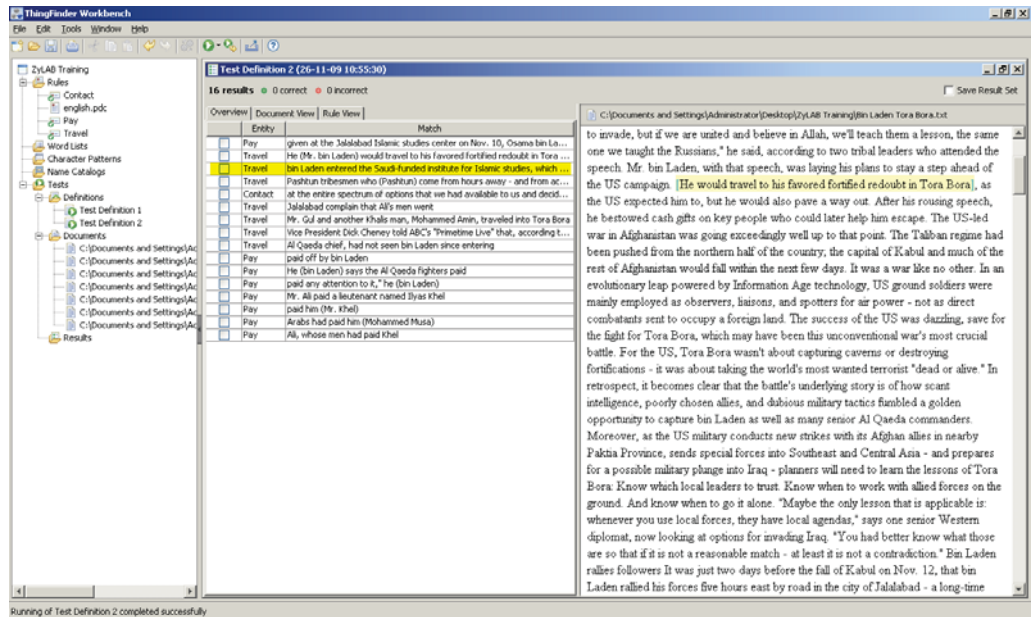
Figure 5: Example of Pronoun Resolution: (He = resolved as Mr, Bin Laden)

Faceted Search and Information Visualization

Text analysis is often mentioned in the same sentence with faceted search and information visualization in large part because visualization is one of the viable technical tools for information analysis after unstructured information has been structured. ZyLAB software extracts facts, entities, and events from data and presents them in logical color-code reports (see figure 6 on the next page).

Another visualization approach is a "treemap," in which an archive is presented as a colored grid (see figure 7 on the next page). The components of the grid are colored-coded and sized based their interrelationships and content volume. This structure allows you to get a quick visual representation of areas with the most entities. A value can also be allocated to a certain type of entity, such as the size of an email or a file.

These types of visualization techniques are ideal for allowing an easy insight into large email collections. Alongside the structure that text analysis techniques can deliver, use can also be derived from the available attributes such as "sender," "recipient," "subject," "date," etc.

The proposed merger between Mega, Inc and CNA Systems, Incorporated, has been postponed, Mega CEO Joe Smith said in an analyst call. "CNA's 1st quarter revenue dropped by 32%, and they lost 23 million dollars," Smith explained. CNA Systems sources blame weak sales in China. CNA shares {CNAI} fell 47 percent to $9.84 on May 12, the first trading day after the announcement.

Company	Mega, Inc
	CNA Systems, Incorporated
Date	May 12
Person	Joe Smith
Person Position	Mega CEO
Currency	23 million dollars
	$9.84
Measurement	32%
	47 percent
Country	China
Noun Group	proposed merger
	analyst call
	1st quarter revenue
	weak sales
	first trading day

Figure 6: Advanced text analytics and classification by ZyLAB

Figure 7: TreeMap visualization of a tree structure

Faceted search, also called faceted navigation or faceted browsing, is a technique for accessing a collection of information represented using a faceted classification, allowing users to explore by filtering available information. A faceted classification system allows the assignment of multiple classifications to an object, enabling the classifications to be ordered in multiple ways, rather than in a single, pre-determined, taxonomic order. Each facet typically corresponds to the possible values of a property common to a set of digital objects.

Facets are often derived by analysis of the text of an item using entity extraction techniques or from pre-existing fields in the database such as author, descriptor, language, and format. This approach permits existing web-pages, product descriptions or articles to have this extra metadata extracted and presented as a navigation facet.

Exploratory search is a specialization of information exploration which represents the activities carried out by searchers who are either:

- unfamiliar with the domain of their goal (i.e. need to learn about the topic in order to understand how to achieve their goal)

- unsure about the ways to achieve their goals (either the technology or the process) or even unsure about their goals in the first place.

Consequently, exploratory search covers a broader class of activities than typical information retrieval, such as investigating, evaluating, comparing, and synthesizing, where new information is sought in a defined conceptual area; exploratory data analysis is another example of an information exploration activity. Typically, therefore, such users generally combine querying and browsing strategies to foster learning and investigation.
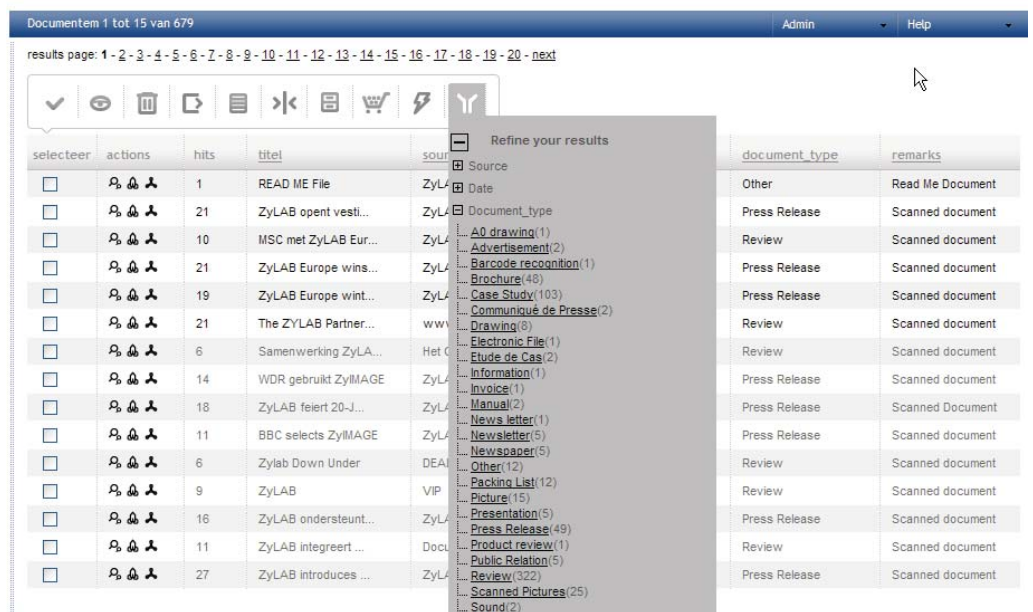
Figure 8: Multiple facets (source, date and document type) are derived automatically from the text by using text analytics. Users see these facets with the individual entity values plus their frequency. By clicking, users can select documents per facet values
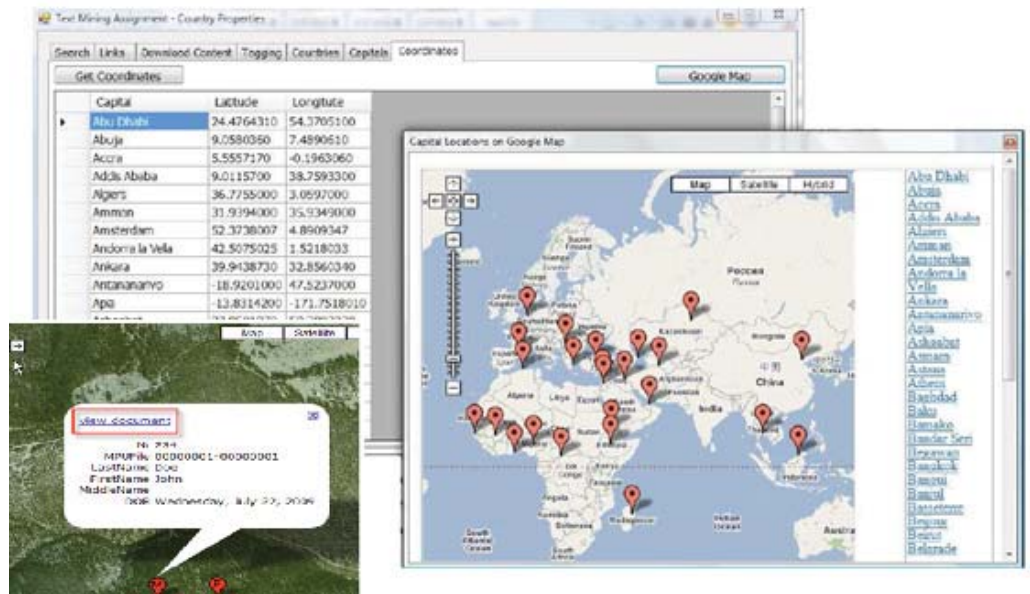
Figure 9: ZyLAB search results can be plotted on a Google Map to show their geographic locations. The image displays the document properties selected by the user and provides a link to open the actual document

Text Analysis on Non-English Documents

As mentioned earlier, many language dependencies need to be addressed when text- analysis technology is applied to non-English languages.

First, basic low-level character encoding differences can have a huge impact on the general search ability of data: where English is often represented in basic ASCII, ANSI or UTF-8, foreign languages can use a variety of different code-pages and UNICODE (UTF-16), all of which map characters differently. Before a particular language's archive can be full-text indexed and processed, a 100% matching character mapping process must be performed. Because this process may change from file to file, and may also be different for different electronic file formats, this exercise can be significant and labor intensive. In fact, words that contain such language-specific special characters as ñ, Æ, ç, or ß (and there are hundreds more of such characters) will not be recognized at all.

Next, the language needs to be recognized and the files need to be tagged with the proper language identifications. For electronic files that contain text that is derived from an optical character recognition (OCR) process or for data that needs to be OCRed, this process can be extra complicated.

Straightforward text-analysis applications use regular expressions, dictionaries (of entities) or simple statistics (often Bayesian or hidden Markov models) that all depend heavily on knowledge of the underlying language. For instance, many regular expressions use US phone number or US postal address conventions, and these structures will not work in other countries or in other languages. Also, regular expressions used by

text analysis software often presume words that start with capitals to be named entities, which is not the case with German. Another example is the fact that in languages such as German and Dutch, words can be concatenated to new words, which is never anticipated by English text analysis tools. More examples of linguistic structures exist that cannot be handled by many US-developed text analysis tools.

In order to recognize the start and end of named entities and to resolve anaphora and co-references, more advanced text analysis approaches tag words in sentences with "part-of- speech" techniques. These natural language processing techniques depend completely on lexicons and on morphological, statistical and grammatical knowledge of the underlying language. Without extensive knowledge of a particular language, none of the developed text analysis tools will work at all.

A few text analysis and text-analytics solutions exist that provide real coverage for languages other than English. Due to large investments by the US government, languages such as Arabic, Farsi, Urdu, Somali, Chinese and Russian are often well covered, but German, Spanish, French, Dutch and Scandinavian languages are almost always not fully supported. These limitations need to be taken into account when applying text analysis technology in international cases.

Content Analytics on Multimedia files: Audio Search

Written text, such as transcripts from audio recordings, cannot fully convey intent, nuance or emotions which are only discernable by human listeners. Additionally, speech-to-text technology is generally limited to dictionary entries.

In contrast, the Audio Search technology of ZyLAB transforms audio recordings into a phonetic representation of the way in which words are pronounced so that investigators can search for dictionary terms, but also proper names, company names, or brands without the need to "re-ingest" the data.

With Audio Search investigators can quickly identify relevant audio clips from multimedia files and from ubiquitous business tools such as fixed-line telephone, VOIP, mobile, and specialist platforms like Skype or MSN Live. The intuitive software enables technical and non-technical users involved in legal disputes, forensics, law enforcement, and lawful data interception to search, review and analyze audio data with the same ease as more traditional forms of Electronically Stored Information (ESI).
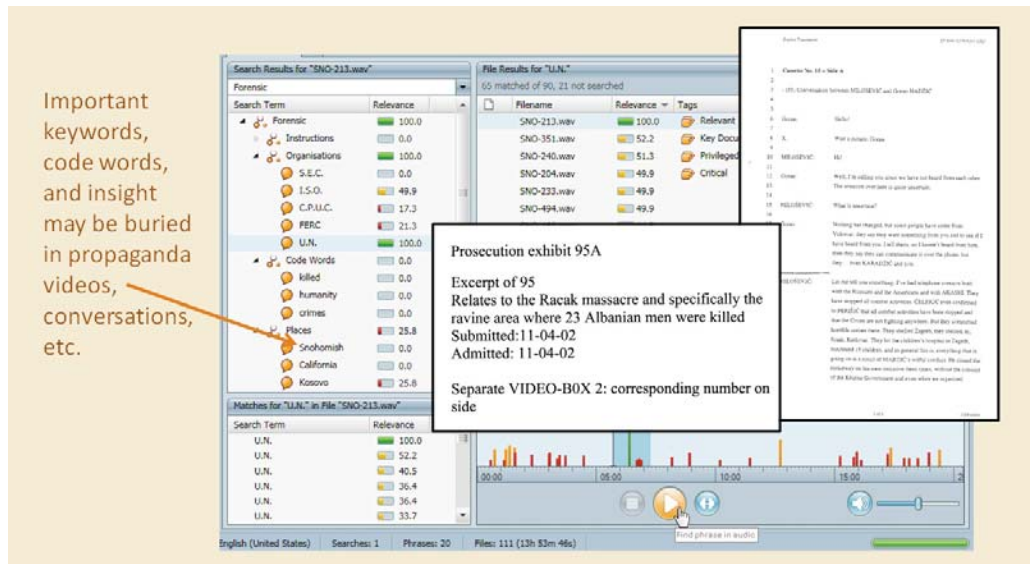
Figure 10: ZyLAB Audio Search shows important keywords, code words and insight that may be buried in recorded conversations, meeting videos and voicemail

A Prosperous Future for Text Analysis

Even with some of the limitations and challenges profiled here, we already see the extensive application of data mining in two areas: e-discovery and compliance. Associated with these are the cognate areas of bankruptcy settlements, due-diligence processes and the handling of data rooms during a takeover or a merger.

The final application in this context will unfold as major legislative changes and stricter control systems will undoubtedly take place in the short term: companies will have to carry out regular (real time) internal preventative investigations, deeper audits and risk analyses. Text analysis technology will become an essential tool to help process and analyze the enormous amount of information in a timely fashion.

Although changes in the legal and financial world are typically evolutions rather than revolutions, a significant role for text analysis in e-discovery and e-disclosure certainly exists. Data collections are just getting too large to be reviewed sequentially, and collections need to be pre-organized and pre-analyzed. With text analysis, reviews can be implemented more efficiently and deadlines can be made easier. The challenge will be to convince courts and auditors of the correctness of these new tools.

Examples of applications are automatic redaction, machine assisted document review and other future legal and investigative power tools.

Figure 11: example of redaction patterns based on text-mining technology

Therefore, a hybrid approach is recommended in which computers make the initial selection and classification of documents and, based on established investigation directives, human reviewers and investigators implement quality control and valuate the investigation suggestions. By doing so, computers can focus on recall, and users can focus on precision.

About ZyLAB

ZyLAB's industry-leading, modular e-Discovery and enterprise information management solutions enable organizations to manage boundless amounts of enterprise data in any format and language, to mitigate risk, reduce costs, investigate matters and elicit business productivity and intelligence.

For nearly 30 years ZyLAB has been a dominant player in compliance and e-Discovery-related solutions, due in part to its' advanced capabilities for multi language support, searching, content analytics, document reviewing, and e-mail and records management (for both scanned and electronic documents).

While the ZyLAB eDiscovery & Production system is generally implemented to investigate a specific legal matter, it is a solid and robust foundation from which to pursue proactive, enterprise-wide objectives for information management. Those broader goals are achieved through the use of the ZyLAB Compliance & Litigation Readiness system.

The ZyLAB eDiscovery system is directly aligned with the Electronic Discovery Reference Model (EDRM) and features modules for forensic sound collection, culling, advanced e-mail conversion (Exchange and Lotus Notes) and legal review.

The company's products and services are used on an enterprise level by corporations, government agencies, courts, and law firms, as well as on specific projects for legal services, auditing, and accounting providers. ZyLAB systems are also available in a Software-as-a-Services (SaaS) model.

ZyLAB's products are extremely open and scalable, with installations managing some of the largest collections of mission-critical data in the world. The award-winning ZyLAB Information Management Platform bundles our core capabilities into a single solution that provides an optimal framework for six, specialized, all-in-one system deployments.
Currently the company has sold 1.7 million user licenses through more than 9,000 installations. All of our solutions include full installation, project management and integration services. Current customers include The White House, Amtrak and US Army OIGs, US Department of Treasury, The EPA, National Agriculture Library, and Royal Library of the Netherlands, FBI, Arkansas and Ohio state police forces, German customs police, and Danish national police, War Crimes Tribunals for Rwanda, Cambodia, and the former Yugoslavia, KPMG, PricewaterhouseCoopers, and Deloitte, Akzo Nobel, Sara Lee, Pacific Life, Siemens, Dow Automotive and Lloyds of London.

ZyLAB has received numerous industry accolades and is one of the few companies to be positioned as a Leader in Gartner's "Magic Quadrant for Information Access Technology" for 2007, 2008 and 2009. ZyLAB has received numerous industry accolades and is one of the few companies to be positioned as a Leader in Gartner's "Magic Quadrant for Information Access Technology" for 2007, 2008 and 2009. In addition, Gartner has given ZyLAB the highest rating ("Strong Positive") in its "MarketScope for E-Discovery and Litigation Support Vendors" for 2007, 2008 and 2009, as well as a "Visionary" rating in its 2011 "Magic Quadrant for E-Discovery".

ZyLAB is certified and registered as compliant with the International Standards Organization (ISO) 9001:2000. ZyLAB also lets Microsoft, Oracle and other infrastructure providers regularly certify critical components that work closely with their infrastructure. ZyLAB was certified under the US-DoD 5015.2 records management standard and ZyLAB is compliant with the European MoReq2 standard and various other regulations

Headquartered in Amsterdam, the Netherlands and McLean, Virginia, ZyLAB also serves local markets from regional offices in New York, Barcelona, Frankfurt, London, Paris, and Singapore. To learn more about ZyLAB visit www.zylab.com

www.zylab.com