Analytics

Big Data – Using Analytics to Improve Your Business

2012 edition

aiim®
The Global Community of Information Professionals
1100 Wayne Avenue, Suite 1100
Silver Spring, MD 20910
Tel 301.587.8202 / 800.477.2446
www.aiim.org

How Content-Analytics can help Big Data

by Johannes C. Scholtes, Chief Strategy Officer, ZyLAB and professor at Text-Mining at the University of Maastricht.

Introduction

The ongoing information explosion from the computer age gained significant momentum in the last decade (or so), finally reaching epic proportions and earning its own name: Big Data. The realities of Big Data encompass both Big Data challenges and opportunities. The challenges stem from the requirements for eDiscovery, governance, compliance, privacy and storage. But the silver-lining to those obstacles is the opportunity to use the collective Big Data to predict and recognize patterns and behavior and to increase revenue and optimize business processes. New data formats (multimedia, in particular), different languages, cloud and other off-side locations and the continual increase in regulations and legislation—which may contradict previous protocols—add even more complexity to this puzzle.

Several commercial tools exist to gain direct access to all such data formats and repositories, regardless of whether they are on-site or off-site and despite the language in which they are composed. But once you have access to the data and you have applied the traditional tools to unpack compound objects (eg. ZIP's and PST's) and manage the vast volumes, how can you derive true understanding from it?

This is where content analytics come into play, and they are becoming an essential toolset, particularly for overcoming the challenges from unstructured and multimedia data. In essence, we need computers to battle the data explosion we've caused with other computers. Applying content analytics helps to assuage the risks of Big Data, but also to benefit from the power of Big Data: broad analysis which yields absolute insights.

Content analytics such as text-mining and machine learning technology from the field of artificial intelligence can be used very effectively to manage Big Data. Think of tasks such as, but not limited to, identifying exact and near-duplicates, structuring and enriching the content of text and multimedia data, identifying relevant (semantic) information, facts and events, and ultimately, predicting what is about to happen or classify information automatically. As a result of these content analytics efforts, users can explore and understand repositories of Big Data better and also apply combinations of advanced search and data visualization techniques easier.

Using these types of automated processes also requires an unbiased evaluation of the results and defensible processes. In other words, the quality and reliability of the automatic structuring, enrichment, classification and prediction techniques needs to be measured by using existing best-practices. Only then will end-users accept the usage of such technology for mission-critical processes. Many such best practices exist from the field of information retrieval.

In this feature article, we will briefly discuss not only content-analytics methods, but also how to evaluate their overall quality and implement them into a defensible process to control the risks of Big Data and to benefit from the predictive power and new insights that can be gained from Big Data.

Content Analytics

Finding Without Knowing Exactly What to Look For

Content analytics such as text-mining, differs from traditional search in that, whereas search requires a user to know what he or she is looking for, text mining attempts to discover information in a pattern that is not known beforehand.

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 as opposed to searching for words, we can search for linguistic word patterns which enable a much higher level of investigation.

In general, text analysis refers to the process of extracting interesting and significant information and knowledge from unstructured text. Text analysis attempts to discover such information through the use of advanced information extraction techniques as well as machine learning. 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.

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. Also, it is hard to find information if you do not know exactly what words are used or if someone does not want to be found. 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 search and the advanced tools needed to control, manage and analyze Big Data.

In addition, regular search does not provide any mechanism to identify, predict or classify specific patterns, which could provide valuable insights into Big Data. This is where text analysis can make a big difference as it is particularly interesting in areas where users must discover new and unknown insights from Big Data.

This is achieved by enriching the original data with additional meta information that allows for not only more sophisticated search capabilities, but also for different context-specific functionalities such as sorting, organizing, categorizing, classifying, grouping and otherwise structuring data based on additional meta-information. In addition, utilizing this additional meta-information will open a whole spectrum of additional search techniques, such as clustering, visualization, advanced (semantic) relevance ranking, automatic document grouping, predicting patterns and categorization.

Know What you are Looking for: What to Investigate

Before you start analyzing Big Data ask yourself what kind of patterns and insights you are interested in as this is essential to defining the specific content analytics, search and data visualization techniques you will need. For example, are you interested in hierarchical, correlational, geographic or temporal patterns? Given the different nature of these analyses, it is virtually impossible to capture them in one analysis, one interface or in one visualization method.

Once your interests are outlined, you can define which information you need to extract, visualize and analyze it and what tools are best to use.

Different Levels of Semantic Information Extraction

One way to structure unstructured data, is to identify and extract relevant meta-information. The options vary from simple extraction methods such as file and document property extraction, to more advanced text analysis options to find semantic information and complex patterns and relationships. Information extraction techniques are often grouped as follows:

- **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 organization.
- **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.

For now, we will focus on information extraction techniques used for Entity, Fact, Attribute, Event and Sentiment as these are the most interesting. Most of the other information mentioned in the list mentioned above can be obtained with standard computer science techniques, but they should not be forgotten as they are the low hanging fruit in content analytics and they can dramatically enhance the search experience and data insights.

How does Information Extraction and Information Enrichment Work?

One of the methods to identify named entities is with the help of regular expressions, which allow data, telephone numbers, Internet addresses, bank account numbers, and social security numbers to be fairly accurately identified. A good example of a regular expression to find an e-mail address is:

\b[A-Z0-9._%+-]+@[A-Z0-9.-]+\.[A-Z]{2,4}\b

More on the exact meaning of the symbols in and structure of this regular expression can be found here: **http://en.wikipedia.org/wiki/Regular_expression**.

Regular expressions can become quite complex, and also quite a lot of work, to define these types of regular expressions, especially because there are many pattern variations; it is not always possible to cover everything with one simple regular expression: either very complex patterns are created, or there are always entities that, due to many irregularities, cannot be identified using regular expressions, even though many templates of regular expressions are widely available as open source.

Another logical approach is to compare the longest occurring form of a named entity with known named entities in dictionaries that contain up-to-date information about what type of entity a specific word or word-combination is. For example: "State University of New York", is that one entity or an entity and a location. Another example is "Mr. and Mrs. Jones", one or two entities? In case of ambiguity (words occurring in multiple dictionaries), linguistic probabilities can be used to select the most probable meaning for a particular entity. For example: if an entity is both a location and a person's name such as in "Mr. Holland", linguistic probability leads immediately to the conclusions that it refers to a person and not to a location.

Explicit rules can also be used to identify named entities. For example, a surname follows a title such as Mr., Mister, Mrs., or Dr. Various relationships between words and their contexts can be formally fixed in this manner. These rules can be created manually, but they can also be derived automatically from large corpora consisting of annotated sample texts.

Rules can also be useful in the recognition of facts (relationships between entities and their attributes).

Examples of extracted named entities, but also of facts and sentiments are shown in figure 1.

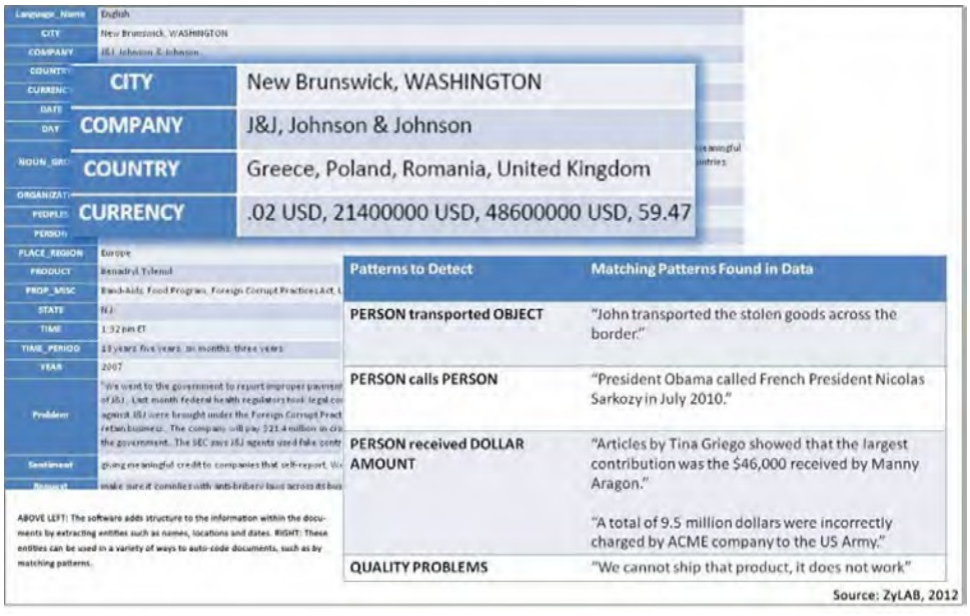

Figure 1: Examples of Extracted Entities, Attributes, Facts and Events

There are some special issues that one has to take into account when applying entity extraction techniques in text mining:

1. Variant Identification and Grouping: It is sometimes necessary to recognize variant names as different forms of the same entity to yield accurate entity counts and records of the location of all appearances of a given entity. For example, one may need to recognize that the word “Smith” in one instance refers to the “Joe Smith” identified earlier so that they can be grouped together as aliases of the same entity.
2. Normalization: Normalizes entities such as dates, currencies, and measurements into standard formats, taking the guesswork out of the metadata creation, search, data mining, and link analysis processes.
3. Entity Boundary Detection: For example, does the text “Mr. and Ms. John Jones” refer to one entity or two? And in the case of “VP John M.P. Kaplan-Jones, Ph.D. M.D.”, where does the entity name begin and end?

In these cases, a basic linguistic analysis (often a Part-of-Speech or POS tagger) is used to help to resolve the ambiguities.

Faceted Search and Information Visualization

Subsequent to the preceding processes, the original data is furnished with diverse supplementary meta properties. At this time, a new, broader range of analysis and search techniques can be applied.

Using the enriched data, derived by the approached discussed here for, it is now possible to organize and visualize data, make complex statistical analyses, call-up similar documents when searching, search on specific features, cluster on attributes, navigate using the complete text of a document and using the available document attributes, etc.

Faceted search is one of these advanced search techniques, also called faceted navigation or faceted browsing. Faceted search uses a collection of information that is represented using a faceted classification, allowing users to explore by filtering available information. Examples of facets are location, color, shoe size, date, author, affiliation, company or other known meta-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 using entity extraction techniques or from pre- existing fields in the database such as author, descriptor, language, and format. This approach permits existing data to have this extra metadata extracted and presented as a navigation facet. See Figure 2 for an example of faceted search.

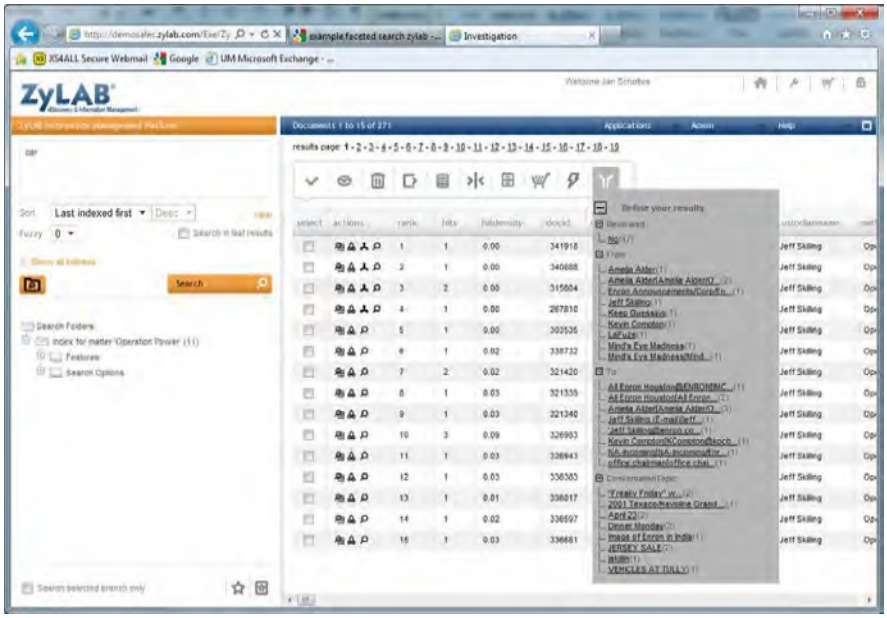

Figure 2: The grey box above provides an example of automatically derived faceted search to refine results on the ENRON data set.

There is more: additional extracted information also allows advanced data visualization such as a star tree or a geographical mapping, as shown in figures 3 and 4.

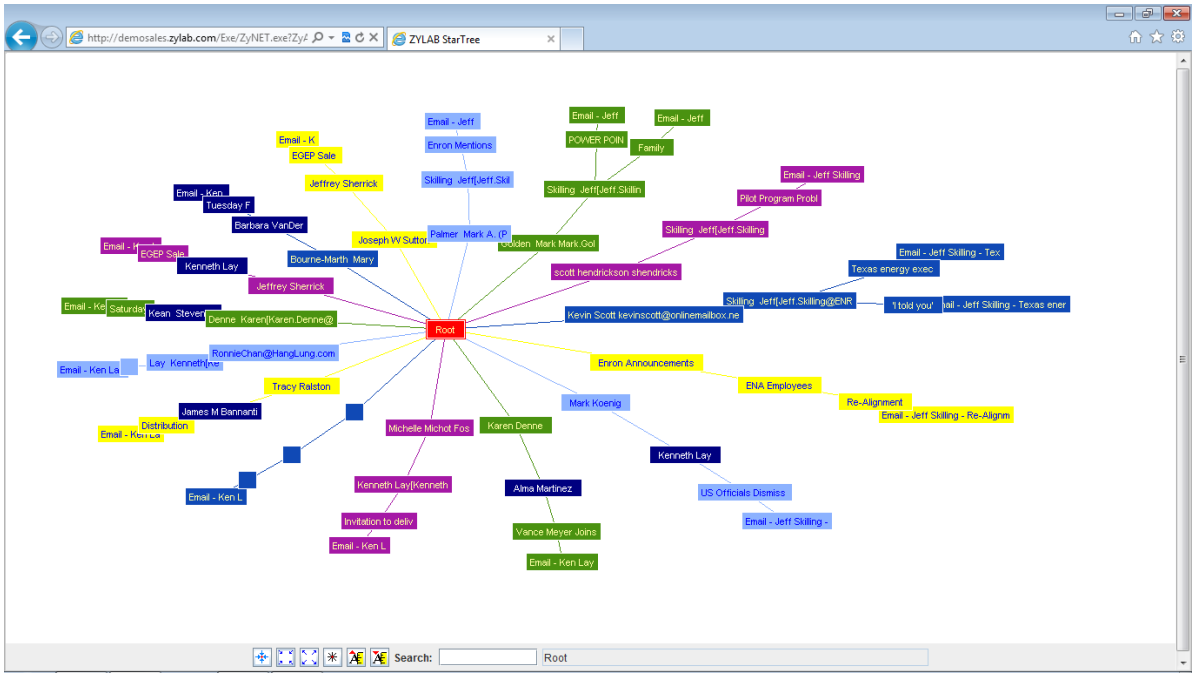

Figure 3: Example of Hierarchical Data Visualization

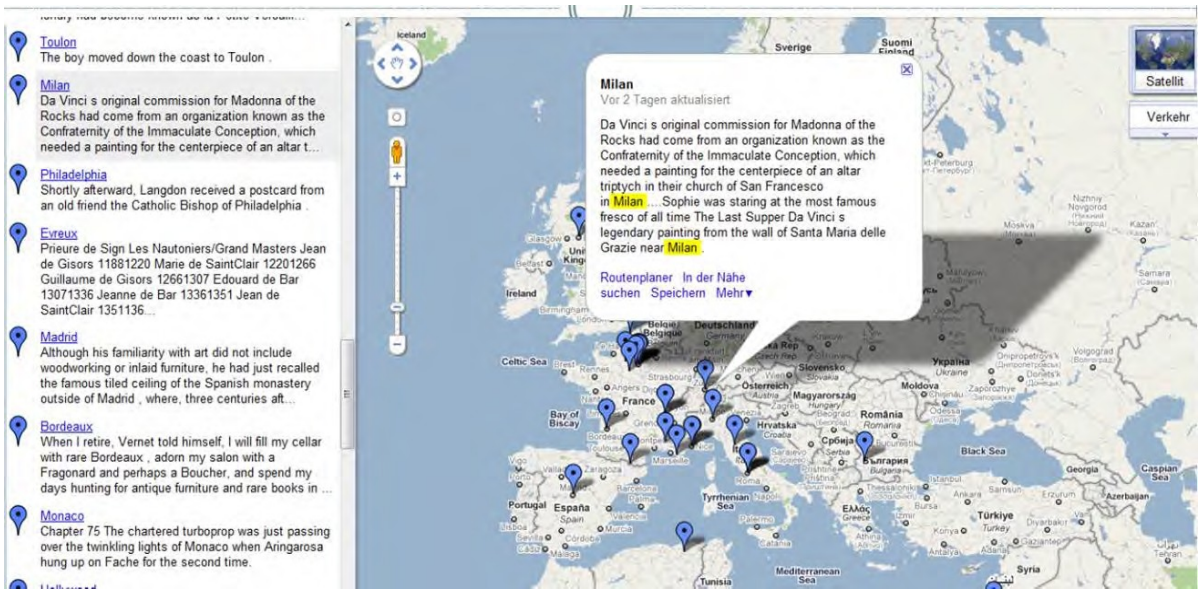

Figure 4: Example of Geographical Mapping where Geo-locations are Derived by Content Analytics

Machine Learning for Prediction: the Holy Grail in data analytics

Both supervised and unsupervised machine learning techniques can be used to predict patterns and reveal more complex insights into Big Data. A machine learning model can be trained with a seed set of documents (samples), which are often annotated documents for particular information categories or known information patterns. Based on these training documents, a machine learning algorithm can derive a model that can classify other documents into the thought classes, or temporal, geographical, correlational or hierarchical patterns can be identified from these training examples.

When applying machine learning, it is wise to first become informed of the potential risks of using the technology for a particular case. It may be the right choice for some cases, but not for others.

Machine learning and other artificial intelligence techniques to predict patterns and behavior are not based on "hocus pocus": they are based on solid mathematical and statistical frameworks in combination with common-sense or biology-inspired heuristics. However, one has to be aware not all of such algorithms are the same, or at least to say, as good. In the case of text-mining, there is an extra complication: the content of textual documents has to be translated, so to speak, into numbers (probabilities, mathematical notions such as vectors, etc.) that machine learning algorithms can interpret. The choices that are made during this translation can highly influence the results of the machine learning algorithms.

For instance, the "bag-of-words" approach used by some products has several limitations that may result in having completely different documents ending up in the exact same vector for machine learning and having documents with the same meaning ending up as completely different vectors. See **http://www.aiim.org/community/blogs/expert/Language-is-Not-Just-a-Jumbled-Bag-of-Words-Why-Natural-Language-Processing-Makes-a-Difference-in-Content-Analytics** for more information on this topic. **The garbage-in, garbage-out principle definitely applies here!**

Other complications arise when:

- More than one foreign language is used in the document set, for instance, if some documents are in English and some documents are in Dutch. Multi-lingual documents in which multiple languages appear in individual documents causes even more problems.
- The more document categories there are, the lower the quality will be for the document classification. This is very logical as it is easier to differentiate only black from white than it is to differentiate 1,000 types of gray values.
- The absence of sufficient relevant training documents will lower the quality of classification. The number of required training documents grows faster than the increase of the number of categorization classes. So, for 2 times more classes one may need 4 times more training documents.
- The documents use very different or very ambiguous language for the same topics (e.g. there are many synonyms and homonyms).
- Dealing with incremental document collections (e.g. new documents are added after training) will result in lower quality or require completely new training of the machine learning.

Several risk factors are listed here, but there are many more depending on the specific machine learning technology that is used. Technology that is based on Bayes classifiers (falsely) presumes statistical independence between words, word occurrences' and other textual properties in the case of analyzing textual data. In layman terms, this means that Bayes classifiers presume that the occurrence of words in text is completely random and that there are no syntactic relations whatsoever. Latent Semantic Indexing (LSI) and its variants such as Probabilistic Latent Semantic Analysis (PLSA) use a lossy information compression algorithm (SVD) that may result in more (irreversible) information loss than required. Knowledge of the specific parameter settings is integral to gaining a full understanding of the quality of specific machine learning models.

If it is not possible or too risky to apply machine learning techniques, then there are also other forms of automatic document classification, such as rules-based document classification. These may be a better choice to use in certain cases, especially when defensibility is an issue. They come with their fair share of set-up, but in almost all cases they are more defensible and easier to manage.

How to evaluate?

There is No Free Lunch

Machine-learning requires significant set-up involving training and testing the quality of the classification model (aka the classifier) , which is a time consuming and demanding task that requires at least the manual tagging and evaluation of both the training and the test set by more than one party (in order to prevent biased opinions). Testing has to be done according to best practice standards used in the information retrieval community (e.g. see the proceedings of the TREC conferences organized by the NIST). Deviation from such standards will be challenged in courts. This is time consuming and expensive and should be factored into the cost-benefit analysis for the approach.

If the classifier does not work (e.g. a mutually-agreed upon predefined quality level is not reached), only retraining the entire model with better training examples will work. Eventually, this process could negate any performance increases or cost savings that could have been achieved by applying the technology. **In that event it is impossible to improve the model and all training and test efforts will have been a waste of time.** This may very well happen in cases that suffer from the complications as described above.

Additionally, one has to be able to explain and defend the application of machine learning technology in court. This may not be a trivial task given the fact that machine learning is based on state-of-the-art principles in linear algebra and probability calculus that are not commonly understood by those who may be involved in the law suit. Therefore, parties and the court will rely heavily on (expensive) expert witnesses.

Content Analytics on Non-English Documents

Due to the global reach of many investigations, a lot of interest also exists for text analysis in multi-language collections. Multi-lingual 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.

Many language dependencies need to be addressed when text analysis technology is applied to non-English content.

First, basic low-level character encoding differences can have a huge impact on the general searchability of data. Whereas 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 an archive with foreign language content 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 various 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 like them) will not be recognized at all if the wrong language model is chosen or if the language is not known.

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 additionally complex.

Straightforward text-analysis applications use regular expressions, dictionaries (of entities) or simple statistics (often Bayesian or hidden Markov models) that all depend heavily upon knowledge of the underlying language. For instance, many regular expressions use US phone number or US postal address conventions (XXX-XXX-XXXX and XXXXX-XXXX, respectively), 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, almost all words can be concatenated to new words, which is never anticipated by English text analysis tools. For instance, the translation for computer mouse cord is the one word vocabulary entry computermuiskabel in Dutch. There are many other linguistic structures 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 upon lexicons and 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 extensive use of this technology by various organizations within 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.

Conclusion

In the last decade, content analytics have been applied with enormous success in fields such as intelligence, security and law enforcement. Sentiment mining (voice of the customer), analytics on clinical research and early detection of product guarantee problems have saved companies millions. Recently, the application of content analytics and machine learning in governance, compliance and eDiscovery have caused technology disruptions in these industries and lead to incredible efficiency and productivity improvements. There is no reason why the application of such techniques cannot make a difference for your business as well.

Even with some of the limitations and challenges profiled here, we already see the extensive application of Big Data analysis not only to manage and control data, but also to benefit from its predictive power and the ability to gain new insights.

One of the expressions in the machine learning community is that “the only good data is more data”. This is essentially what Big Data is all about. Big Data will help us to develop better content analytics and better machine learning. Content analytics will help us better to control, manage, understand and use Big Data.

As a result, one can state that Content Analytics and Big Data are symbiotic: one cannot only exist without the other, but they will also enhance one another to become better and more advanced in the very near future!

Johannes C. Scholtes, Chief Strategy Officer, ZyLAB and professor at Text-Mining at the University of Maastricht.

Johannes C. Scholtes is Chief Strategy Officer of ZyLAB. Scholtes was the President / CEO of ZyLAB from 1987 to 2009, at which time he recruited a long-term business associate to assume those duties so that Scholtes could focus 100% on weaving his background in knowledge engineering through the ZyLAB strategy. Scholtes has been involved in deploying in-house eDiscovery and investigative software with organizations such as the United Nations War Crimes Tribunals, the FBI-ENRON investigations, the Executive Office of the President of the United States of America, and thousands of other users worldwide.

Scholtes holds the Extraordinary Chair in Text Mining from the Department of Knowledge Engineering at the University of Maastricht in The Netherlands, and he is a member of the AIIM board: the worldwide association for information management.

Before joining ZyLAB, Scholtes was lieutenant in the intelligence department of the Royal Dutch Navy. Scholtes holds a M.Sc. degree in Computer Science from Delft University of Technology and a Ph.D. in Computational Linguistics from the University of Amsterdam.

1100 Wayne Avenue, Suite 1100
Silver Spring, MD 20910
Tel 301.587.8202 / 800.477.2446
www.aiim.org