INTERNATIONAL INSTITUTE FOR ANALYTICS

Enterprise Research Service

Research Brief

Text Mining Versus Text Analytics

August 2014

Part II of II

Authored by:

Andrew Fast, Ph.D. , Chief Scientist, Elder Research, Inc.

John F. Elder, IV, Ph.D. , Founder and CEO, Elder Research, Inc.

Copyright©2014 International Institute for Analytics. Proprietary to subscribers. IIA research is intended for IIA members only and should not be distributed without permission from IIA. All inquiries should be directed to membership@iianalytics.com.

Key Takeaways

1. Text Mining and Text Analytics are complementary ways to automatically extract meaning from text. They solve the same problems, but use different techniques.
2. Of the two, text Analytics is the older discipline and developed within the field of computational linguistics. Its strength is the ability to encode human understanding into a series of linguistic rules. Rules generated by humans are high in *precision*[1], but they do not automatically adapt and are usually fragile when tried in new situations.
3. Text mining is a newer discipline arising out of the fields of statistics, data mining and machine learning. Its strength is the ability to inductively create models from collections of historical data. Because statistical models are learned from training data they are adaptive and can identify "unknown unknowns", leading to better *recall*. Still, they can be prone to missing something that would seem obvious to a human.
4. In our experience and from historical comparisons, text analytics and text mining approaches have essentially equivalent performance. However, the type of work the analyst performs to achieve those results differs dramatically. Text analytics requires an expert linguist to produce complex rule sets, whereas text mining requires the analyst to hand-label cases with outcomes or classes to create training data.
5. Due to their different perspectives and strengths, combining text analytics with text mining often leads to better performance than either approach alone.

Introduction

We define *textual analysis* to be the automated analysis of unstructured textual data, containing within it the methodologies of text mining and text analytics. Leading textual analysis use cases include Sentiment Analysis, Natural Language Processing (NLP), Information Extraction, and Document Categorization. Historically, text analytics practitioners have backgrounds in computational linguistics and knowledge management, whereas text mining practitioners come from the fields of data mining and statistics.

[1] A search method's *Precision* is measured by the proportion of retrieved documents that are relevant. Its *Recall* is the proportion of relevant documents that are retrieved. Precision = (true positives) / (true positives + false positives). Recall = (true positives) / (true positives + false negatives). You want both to be high, but there is a trade-off since retrieving more documents, for instance, will raise recall but lower precision.

Copyright©2014 International Institute for Analytics. Proprietary to ARC subscribers. IIA research is intended for IIA members only and should not be distributed without permission from IIA. All inquiries should be directed to membership@iianalytics.com.

Because of the resulting differing aptitudes and mindsets, there is a friendly (and sometimes not so friendly) competition between practitioners of both approaches as to superiority. In general, the linguistic and the statistical approaches can each solve the same types of problems using the same original data. What is most different between them is how *features* – characteristics of the documents – are combined. Text analytics relies on linguistic rules whereas text mining utilizes statistical models. For the linguistic approach, improvements are made by studying details of its performance and coming up with rules that cover exceptions without somehow jeopardizing the performance of earlier rules. Improvements for the statistical approach are obtained by analysts identifying where in data space more example cases are needed, labeling more training data cases, and then fitting a more thorough model.

Consider the problem of *named entity recognition* that is identifying person names from text. Key features for identifying a person name have been found to include two capitalized words in the middle of a sentence (e.g., Bill Gates), the use of an honorific (e.g., Mr., Mrs., Miss), or appearance of a tailored word list (e.g., US Census Name List). These work fairly well immediately, yet none of these rules for identifying person names is absolute; there are always exceptions.

For example, consider the following sentence from Wikipedia on **Mister Donut**, a franchise now based primarily in East Asia[2]:

> `The Mister Donut business became so popular that Winouker and Slater decided to go into franchising.`

Using only the standard features described above, the phrase “Mister Donut” will likely be identified as a person name. Rule-based approaches can avoid this problem with an exception indicating that “Donut” should never be marked as a person name. Alternatively, a rule could include other such contextual information and indicate, for instance, that “the” rarely precedes a person. Statistical approaches assign a probability to whether each feature of a word (e.g., capitalization) indicates a person name and then combine probabilities of features from the term in question and its surrounding terms. By including the previous and following terms (often multiple in each direction) statistical approaches can detect from training data that person names are rarely preceded by “the” or followed by “business” leading to a decreased probability.

The remainder of this brief provides background on the two technology approaches and compares their strengths and weakness. We highlight three different use cases, each exhibiting different positive/negative error trade-offs, and suggest which approach works best. Lastly, we describe how linguistic and statistical approaches can be combined for best success.

---

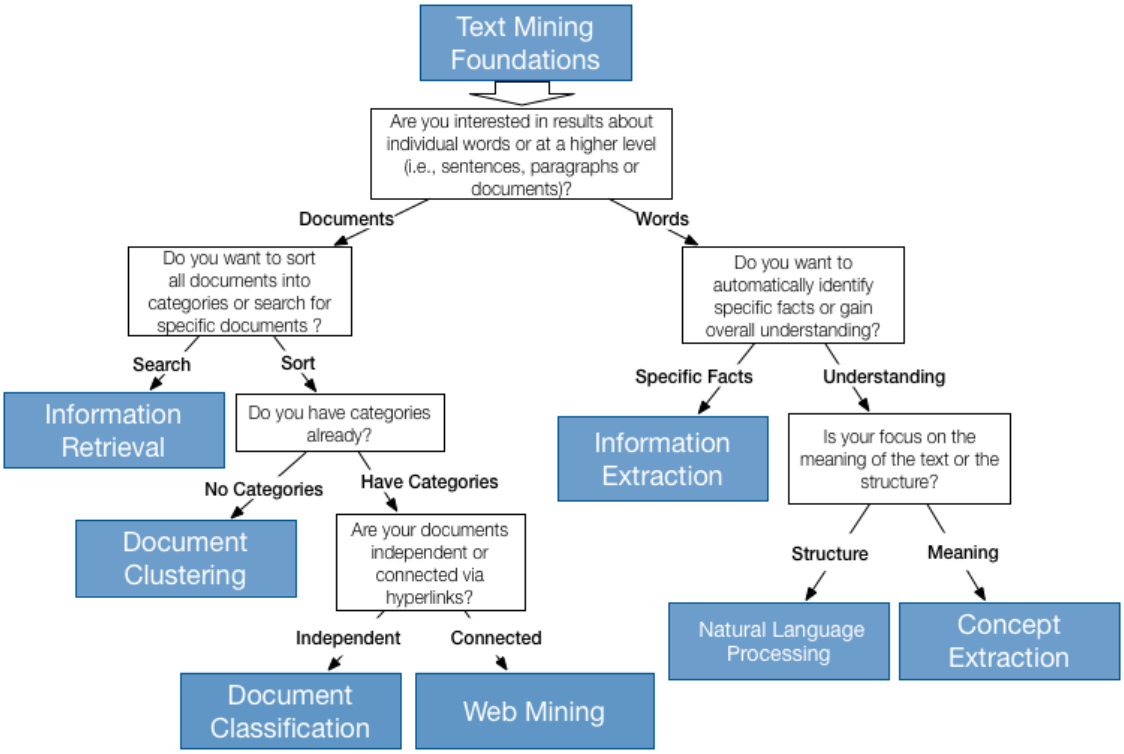
[2] http://http://en.wikipedia.org/wiki/Mister_Donut

Copyright©2014 International Institute for Analytics. Proprietary to ARC subscribers. IIA research is intended for IIA members only and should not be distributed without permission from IIA. All inquiries should be directed to membership@iianalytics.com.

The Problem of Textual Analysis

When someone is interested in performing textual analysis we have found that they could mean one of seven different goals or "practice areas", as illustrated in the tree-shaped diagram of Figure 1[3]. The seven practice areas are the blue terminal nodes or "leaves" of the (upside-down) tree, with an eighth area at the top representing textual analysis foundation topics.

By answering a handful of questions about your goals and the types of your data, and following the resulting path, this diagram identifies your practice area of interest. (The book *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications* further lists the chapters, tutorials, and chief external resources to draw from to succeed at that task.) Your goal has a large effect on which type of technology – linguistic or statistical – to choose, as will be highlighted in the three case studies to follow.



[3] From Chapter 2 of *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications,* G. Miner, D. Delen, J. Elder, A. Fast, T. Hill, and R. Nisbet, Elsevier, January 2012

Copyright©2014 International Institute for Analytics. Proprietary to ARC subscribers. IIA research is intended for IIA members only and should not be distributed without permission from IIA. All inquiries should be directed to membership@iianalytics.com.

Figure 1: A visual breakdown of the different textual analysis tasks or practice areas

The first distinction is whether your units of interest are documents or words. Document-centered processing tasks include keyword search, document classification, and document categorization. Word-based processing includes information extraction goals such as named entity recognition, and natural language processing tasks such as part of speech identification, phrase identification, and grammatical parsing.

Naturally, all textual analysis starts with a collection (or *corpus*) of unstructured text. Often the corpus contains textual snippets from a database or other structured source (e.g., XML), so the data is semi-structured (though here we can treat it as unstructured). Whether focusing on words or documents, facts or understanding, sorting or searching, etc., the linguistic or statistical approaches both use the same building blocks: the characteristics, or *features*, of words contained within corpus documents. Common features include capitalization (useful for identifying proper nouns), inclusion of a period (indicates abbreviations or acronyms), membership within a defined list (lexicons), and a word's relationship with surrounding words.

Major distinctions between text mining and text analytics are summarized in Table 1.

	Text Mining	Both	Text Analytics
Approach	Statistical Machinery		Linguistic Rules
Inputs		Features of words and documents	
Performance		Equivalent (eventually)	
Effort	Labeling training data		Tuning rule-sets

Copyright©2014 International Institute for Analytics. Proprietary to ARC subscribers. IIA research is intended for IIA members only and should not be distributed without permission from IIA. All inquiries should be directed to membership@iianalytics.com.

Strength	Flexibility		Human Understanding
Rule Generator	Algorithmic		Human

Table 1: Text Mining and Text Analytics

Distinctives of Text Analytics

Traditional text analytics approaches focus on the deep linguistic understanding of text. This requires the encoding of human knowledge into structured formats including:

- Taxonomies - hierarchical organization of terms and concepts
- Ontologies and Semantic Networks - graph-structured organization of terms and concepts
- Lexicons - a catalog of terms
- Regular expressions - rules for identifying specific patterns from text

Words from these different structures are combined into linguistic rules that suggest outcomes and labels for documents and parts of text.

The key to most text analytics approach is the human generation of rules. Human experience is usually correct for the straightforward answers (see e.g., name identification), and thus has high precision for the most common cases (i.e., no obvious misses). With more complex concepts, however, it can be difficult to identify the inputs and prioritize the rules correctly to get the right answers leading to lower recall.

The complexity of creating working rules means most of the effort in a text analytics process is generating rules to account for the wide variety of exceptions that are present in a text corpus. Success typically requires an expert practitioner, often with a background in computational linguistics. The end result of a text analytics development process is a large rule-base that is typically quite accurate for the problems for which it was developed, but hard to adapt to changing data or new domains.

Copyright©2014 International Institute for Analytics. Proprietary to ARC subscribers. IIA research is intended for IIA members only and should not be distributed without permission from IIA. All inquiries should be directed to membership@iianalytics.com.

Distinctives of Text Mining

Statistical approaches can use all of the same inputs as text analytics, though often not to the same detail, since, for the most general problems simple statistics often capture the majority of the meaning without one needing to create sophisticated taxonomies or other structured lists.

The power of text mining comes from inductive modeling from labeled training data. Unlike the top-down approach of text analytics, inductive learners make inferences from large numbers of examples, using probabilities to measure the uncertainty from conflicting cases. This approach produces results with high recall, allowing the modeling algorithm to identify weak patterns that would otherwise go unnoticed. Finding possibly surprising documents in this way is one of the strengths of a statistical approach. Probabilistic weights also provide a mathematically sound method for combining evidence from multiple kinds of features. Examples of this approach include Naive Bayes classifiers, Hidden Markov Models and Conditional Random Fields.

One of the key challenges of using a statistical approach is preventing the models from *over fitting* the training data. Models over fit when the training algorithms try too hard to fit the known training data ("memorizing" it, instead of learning the general concepts), leading to lower performance on future (unseen) data. Over fit models are results of sparse training data which makes relationships between features and outcome labels appear stronger than they really are. It is fallout from the "vast search effect" in looking for the best model, or from biases and inefficiencies in the model training process.

But the modeling doesn't need to be too complex. For example, James Pennebaker, the chair of the Department of Psychology at the University of Texas Austin, recently authored *The Secret Life of Pronouns*, a book describing the truly informative nature of *function* words. These ordinary words include pronouns (such as I, you, they), articles (a, an, the), prepositions (to, of, for), auxiliary verbs (is, am, have). Pennebaker shows how simple models of function word usage can be used to detect supervisor/supervisee relationships, guilty vs. innocent defendants from their testimony, and many other examples. This work highlights the strength of text mining: gain insight into the information reflected in text data without the complexity of needing a deep understanding of language and the domain.

Balancing the Strengths of Statistical and Linguistic Approaches

Because text mining and text analytics solve the same problems, use the same inputs, and are often confused for one another by non-experts, it can be tough to know which method is best for

Copyright©2014 International Institute for Analytics. Proprietary to ARC subscribers. IIA research is intended for IIA members only and should not be distributed without permission from IIA. All inquiries should be directed to membership@iianalytics.com.

a given problem. The key criteria for choosing between the two are the cost and effort required to implement a complete solution, and the desired error profile of the results.

The highest profile type of error is a *false positive* error, also known as a type I error or false alarm. This error occurs when the model or rule-base detects an event when none exists. For example, a false positive occurs when text analysis identifies a part of text as relevant or interesting when it is not. The opposite error, known as a *false negative* occurs when the model or rule-base fails to detect an interesting part of text that it should have. False negatives are also known as type II errors or misses.

There are almost always many potentially interesting documents in a corpus, and *precision* and *recall* are the primary metrics for how well a system is identifying them. Precision measures the proportion of documents the system classifies as interesting that truly are to the user. Recall measures the proportion of truly interesting documents found.

It is our experience that statistical systems (text mining) have fewer false negatives; they find more unknown but interesting cases. Conversely, linguistic systems (text analytics) have fewer false positives; when they claim something is interesting it more often is. Consequently, the primary decision point for determining the balance between the methods is the cost of a false positive error relative to the cost of a false negative error. If false positives and otherwise incorrect information are potentially harmful to brand reputation or user adoption, then leaning towards to a linguistic rule-based approach make sense. Conversely, if there is a large cost to missed information and interesting information is hard to define in a concise way then the statistical approach allows for broader discovery of new information.

To better highlight these tradeoffs, we consider three popular use cases of text analytics: Knowledge Management in Established Domains, Document Categorization and Prioritization, and Named Entity Recognition.

Use Case 1: Knowledge Management in Established Domains

Knowledge Management is closely related to search and document categorization. Its goal is to provide documents to users within a known taxonomic structure. Examples include the medical and legal fields where industry standard taxonomies and classifications are well developed.

In these formal applications, the cost of a false positive is quite high as users will be directed to incorrect content, reducing trust in the system. Linguistic rules are used widely here because of their traceability and certainty. Text analytics is primarily used for auto-tagging that is assigning documents to certain branches in the taxonomy. These tagging rules are typically assigned based on the presence of specific terms in the document.

Copyright©2014 International Institute for Analytics. Proprietary to ARC subscribers. IIA research is intended for IIA members only and should not be distributed without permission from IIA. All inquiries should be directed to membership@iianalytics.com.

When no formal taxonomy exists or the domain is changing rapidly, linguistic processing loses its edge as there is no standard assignment of the documents. This can be addressed by the creation of a formal taxonomy, usually a lengthy process, or by the use of statistical approaches to augment the knowledge management system. For example, topic models can suggest groupings of words to create taxonomy-like categories. Similarly, statistical clustering and document similarity algorithms can be employed to group documents dynamically based on shared content without the requirement of a formal taxonomy. Including statistical approaches, though, has the potential to reduce the precision of the results.

Use Case 2: Document Categorization and Prioritization

For many problems documents or snippets of text must be categorized into categories where the rules for categorization are difficult to define. Take, for example, the problem of approving disability claims based on the free text field containing the applicant's specific allegations of multiple health issues. The inputs to the decision process include their current health and the number and severity of symptoms. The determination of health level (needed to assess approval) requires the expert opinion of a trained medical professional, which is difficult and time-consuming to mimic via a series of rules.

Instead, a statistical model can be built to infer the characteristics that indicate disability level based on thousands of past decisions of application adjudicators. Relationships between the features and the judgment induced from many different cases seen by many different adjudicators can be used to develop the best decision model for new cases. This model includes weighting each symptom by its historical probability of leading to an approval, and combining these probabilities in a sensible way to account for applicants with multiple health issues.

To make text mining work, the text must first be converted into a numerical structure that can be input into statistical algorithms. The simplest type of conversion is the "bag of words" strategy where an input feature is created for each individual word: a 1 if the word is present, and a 0 if not. This strategy works quite well despite its simplicity. Still, there are obvious short-comings. For example, multi-word phrases (e.g., Amyotrophic lateral sclerosis) are treated as three different inputs rather than a single output. The best way to counter this problem (and many others) is a targeted use of text analytics to create richer linguistic features that become inputs into the modeling algorithms.

Use Case 3: Named Entity Recognition and Natural Language Processing

Named entity recognition was one of the earliest battlegrounds for the statistical and linguistic approaches. It is a sequence classification task where strings of tokens are identified as being a

Copyright©2014 International Institute for Analytics. Proprietary to ARC subscribers. IIA research is intended for IIA members only and should not be distributed without permission from IIA. All inquiries should be directed to membership@iianalytics.com.

named entity or not. The exact definition of a *named entity* varies from task to task, but the goal is to identify collections of people, places, organizations, or other proper nouns from text.

Over a decade ago, machine researchers recognized that Hidden Markov Models and the then-new Conditional Random Field model could create effective entity recognition algorithms. In light of these discoveries, the Conference for Natural Language Learning sponsored a number of competitions between the new statistical methods and the incumbent rule-based approaches. After multiple competitions the performance of both methods – when used by experienced practitioners – was equivalent.

Now the consensus is that combining the two approaches is the most effective way to perform named entity recognition. Linguistic rules using lexicons, taxonomies, and ontologies are more effective at identifying rare entities with predictable structure such as organization names. There are typically not enough mentions of these types of entities to form good enough training data for statistical algorithms to identify them consistently. But statistical algorithms tend to be more effective at identifying less predictable entities such as person names. As mentioned earlier, there are many useful but imperfect pieces of evidence that must be combined to identify person names. Probabilistic models inferred from data are best able to balance the competing evidence.

Key Points

- ***Text mining and text analytics can each be used be to solve any text analysis problem*** – Choosing the right approach (or mix) depends on whether the problem is well-defined or open-ended, whether there are historical labeled data available or well-established lists of keywords, and the cost of false positive and false negative errors. For rapidly changing domains, statistical approaches are able to identify weaker patterns that are predictive, whereas updating linguistic rules can be very labor intensive. These characteristics lead to a natural precision/recall trade-offs. Statistical approaches have better recall "out of the box", but linguistic rules have higher precision. The best solutions find the right balance given the specific business problem.
- ***Improving text analytics with text mining*** – For text analytics projects, there are a number of ways to incorporate statistical text mining to improve the results. Most pure text analytics practitioners view text mining as a method for exploring the corpus and suggesting possible rules. For example, statistical approaches can quickly identify words with similar meanings and/or usage, identify important keywords, and suggest possible multi-word phrases. This additional information can help guide the creation of new linguistic rules.

Copyright©2014 International Institute for Analytics. Proprietary to ARC subscribers. IIA research is intended for IIA members only and should not be distributed without permission from IIA. All inquiries should be directed to membership@iianalytics.com.

Beyond suggesting new rules, text mining can replace or augment existing linguistic rules. One of the strengths of a statistical approach is the ability to combine evidence from multiple features. As rule-sets increase in size, complexity, and the number of special cases, text mining can reduce the rule maintenance burden and increase the ability to uncover new and surprising knowledge from the corpus.

- ***Improving text mining with text analytics*** – Text mining uses statistical approaches to combine multiple features into a single decision. The best way to improve text mining is to upgrade the quality of the features through traditional text analytics approaches such as lexicons, taxonomies, and rules. These help to ensure that feature creation follows "common sense", including not breaking multi-word phrases, creating domain-specific linguistic rules, and accounting for technical language.
- ***Driven by continued growth in online applications such as targeted advertising, statistical approaches for textual analysis are one of the fastest growing areas of machine learning*** – The truly "big" data associated with most online tasks amplifies the need for the rapid scalability provided by a statistical approach. Look for the rapid expansion of statistical text mining that began with Google in the late 1990s to continue for the foreseeable future.

Conclusion

Linguistic and statistical approaches for processing text provide complementary results for extracting value from unstructured textual data. Though each has been practiced independently, the most effective solutions combine their strengths. This balances the precision of linguistically-based text analytics with the powerful recall of a statistical text mining approach. The rapid growth of "big data" and predictive analytics means that the best techniques for achieving this balance will be constantly evolving, yet the tools exist today to make great progress on the wide variety of textual analytics challenges.

About the Authors

Andrew Fast, Ph.D. is the Chief Scientist at Elder Research and leads the research and development of new tools and algorithms for data and text mining. Dr. Fast graduated Magna Cum Laude from Bethel University and earned Master's and Ph.D. degrees in Computer Science

Copyright©2014 International Institute for Analytics. Proprietary to ARC subscribers. IIA research is intended for IIA members only and should not be distributed without permission from IIA. All inquiries should be directed to membership@iianalytics.com.

from the University of Massachusetts Amherst. There, his research focused on causal data mining and mining complex relational data such as social networks. Dr. Fast has published on an array of applications including detecting securities fraud using the social network among brokers, and understanding the structure of criminal and violent groups. Other publications cover modeling peer-to-peer music file sharing networks, understanding how collective classification works, and predicting playoff success of NFL head coaches (work featured on ESPN.com). With colleague Dr. John Elder and others, Andrew has written a book on *Practical Text Mining* that was awarded the *PROSE* Award for Computer Science in 2012.

John Elder, Ph.D. has authored innovative data mining tools, is a frequent keynote speaker, and was co-chair of the 2009 Knowledge Discovery and Data Mining conference, in Paris. Dr. Elder co-authored 3 books (on practical data mining, ensembles, and text mining), two of which won "book of the year" awards in Mathematics or Computer Science. John's courses on analysis techniques -- taught at dozens of universities, companies, and government labs -- are noted for their clarity and effectiveness. Dr. Elder earned Engineering degrees from Rice and UVA and is an Adjunct Professor of Systems Engineering at UVA. He was honored to be named by President Bush to serve 5 years on a panel to guide technology for national security. Lastly, John is grateful to be a follower of Christ and the father of five.

Elder Research Inc. (ERI) is the US's leading consulting company in data mining, predictive analytics, and text mining. Founded by John Elder, in 1995, ERI has helped government agencies and Fortune Global 500® companies solve real-world problems by amplifying the productivity of their analysts. Drawing from experience in multiple industries, ERI brings cutting-edge technology into front-line practice to achieve high return on investment. Headquartered in Charlottesville, Virginia, ERI also has a growing office in Washington, DC.

Copyright©2014 International Institute for Analytics. Proprietary to ARC subscribers. IIA research is intended for IIA members only and should not be distributed without permission from IIA. All inquiries should be directed to membership@iianalytics.com.