Mission Possible: Connecting Structured and Unstructured Data to Create New Insights

Advanced text analytics can link structured data with unstructured data in ways that were impossible years ago. These capabilities are unlocking insights and enabling new workflows in business domains where entities—people, places, organizations, disease or drug names and more—are the connectors between data sources.

by Tina Lieu and Steve Kearns
ABOUT BASIS TECHNOLOGY

Basis Technology provides software solutions for text analytics, information retrieval, digital forensics, and identity resolution in over forty languages. Our Rosette® linguistics platform is a widely used suite of interoperable components that power search, business intelligence, e-discovery, social media monitoring, financial compliance, and other enterprise applications. Our linguistics team is at the forefront of applied natural language processing using a combination of statistical modeling, expert rules, and corpus-derived data. Our forensics team pioneers better, faster, and cheaper techniques to extract forensic evidence, keeping government and law enforcement ahead of exponential growth of data storage volumes.

Software vendors, content providers, financial institutions, and government agencies worldwide rely on Basis Technology’s solutions for Unicode compliance, language identification, multilingual search, entity extraction, name indexing, and name translation. Our products and services are used by over 250 major firms, including Cisco, EMC, Exalead/Dassault Systems, Hewlett-Packard, Microsoft, Oracle, and Symantec. Our text analysis products are widely used in the U.S. defense and intelligence industry by such firms as CACI, Lockheed Martin, Northrop Grumman, SAIC, and SRI. We are the top provider of multilingual technology to web and e-commerce search engines, including Amazon.com, Bing, Google, and Yahoo!.

1. ABSTRACT

Advanced text analytics can link structured data (databases and spreadsheets) with unstructured data (news stories, social media, webpages, and documents) in ways that were impossible years ago. These capabilities are unlocking insights and enabling new workflows in business domains where entities—people, places, organizations, disease or drug names and more—are the connectors between data sources.

In the past, research in these domains was limited to keyword search and filtering or faceting on structured or semi-structured metadata. Entities in unstructured text were manually identified, and looked up in structured databases. Today, by combining automatic entity extraction with name matching, users can automatically identify entity mentions in unstructured data and link them with structured information. This linkage simplifies the process and combines data about an entity into a complete picture, rather than just returning simple search results. This approach goes beyond general problem-solving to linking specific content or topic domains where entities are disproportionately important.

This paper explores how the combination of these two technologies is revolutionizing the decision-making and problem solving process in industries such as e-discovery, social media monitoring, government intelligence, financial compliance, and publishing.

2. CONNECTING DATA: THE GENERAL CASE

Let’s suppose you are writing an article about Tim Cook, CEO of Apple, Inc., or your company is looking to invest in Apple and is performing due diligence. Entity extraction will comb through unstructured data, finding the entity “Tim Cook” in news articles, and as well as instances where his name might be written “Timothy D. Cook” in SEC filings or “Tim D. Cook.” Then, name matching capabilities allow us to identify these three mentions as being very similar. By linking these mentions with “Apple” the company (rather than the fruit) the system can distinguish him from Tim Cook the football player; Tim Cook the ice hockey defenseman; and Tim Cook the mass communications scholar.

Unstructured data might include news stories or blog posts about Tim Cook. Semi-structured data includes Cook’s LinkedIn profile about his previous positions and schooling or the results of a news service search through LexisNexis. Structured information can be retrieved about Apple Inc., its stock price, and competitors via Yahoo! Finance, or internal structured databases.

Linking structured and unstructured information requires two key technologies: entity extraction and name matching.

Entity extraction adds structure to unstructured data. The example below, a paragraph of the Wikipedia article of Tim Cook, illustrates how the extracted entities create structured metadata about the article, allowing it to be linked to structured data in a database. This way, a publisher can link this article to other documents about “Tim Cook” and the organization, “Apple.”
Name matching is essential as entity names often link disparate data sources. One challenge is that names are often written differently in different sources. Casual references use nicknames or partial references, while formal documents and databases may contain typos, omission/inclusion of a middle name, or variable spelling of foreign names (e.g., Abul-qassem El-Chabbar vs. Abo Al-Qassim Al-Shabbi). Furthermore, in multilingual datasets, the same name can be written in different languages (e.g., Mao Zedong vs. Mao Tse Tung vs. 毛泽东). When linking information across data sets is mission-critical, accurate name matching is indispensable.

Let’s now explore how these two technologies can be leveraged within specific domains:

1. Publishing and information services
2. E-Discovery
3. Financial compliance
4. Social media monitoring and analysis
5. Government intelligence & law enforcement

3. PUBLISHING AND INFORMATION SERVICES
Publishers, such as Factiva or LexisNexis, typically publish semi-structured articles that follow a specific format. Each document has structured metadata such as author, title, publication date, etc., but then the document body will be largely unstructured data. Increasingly, there is a move...
to better leverage the value in this content by making it more discoverable and thus, more useful. Today, additional metadata about entities is often added by a team of skilled editors and content curators, who manually identify references and link them against master databases.

To take advantage of the power of automatic or interactive data linking, these publishers could leverage entity extraction and entity matching to streamline their process.

*Entity extraction* automatically locates the mentions of entities—such as people, places, and organizations—creating structured metadata for each entity. Domain-specific entities will be extracted for specialized fields as science, pharmaceuticals, patent search, or finance.

Precisely locating mentions of entities also allows publishers to link across documents and show users related information more easily, letting them further enrich their data repository for more targeted uses. Alternatively, content in the repository can be organized around entities, rather than documents, providing rich context about a particular person or organization.

*Name matching* helps editors link references of the same entity (Tim Cook, Timothy Cook, or Timothy D. Cook) to their structured, or “master,” entity database, so that editors can spend time on the tougher job of distinguishing between multiple people named “Bob Jones.” Name matching can also be used by end users at query time, to help them find the content they are looking for, no matter how entities appear in the text—casual references, translated names, nicknames, etc.

Data is the primary resource that publishers possess, and its value increases as it becomes easier for users to find what they are looking for. Linking articles against structured databases using entity extraction and name matching can more efficiently build the value of existing content.

**4. E-DISCOVERY**

More and more documents in the discovery phase of a lawsuit are electronic, whether as email, office documents, address books, or text message logs. Items such as email address books are highly structured with fields like name, address, email, phone number, etc. Email messages are semi-structured with the “to” and “from” fields specified, but the body of the email entirely unstructured. Office documents are mostly unstructured except for basic properties such as “author” and “last saved date.”

Entity extraction plays a role here, identifying names from email bodies and documents, while name matching can be used to match these entities against structured entries in cellphone or email address books.
Perhaps the cellphone log of a central figure in the lawsuit reveals that he often called and texted “Bobby Shakelton,” who was not originally seen as related to the case. But then upon looking for “Bobby Shackleton” in other electronic documents, entity search might find evidence in email from Bobby, where he appears as “Robert D. Shackleton.” Bobby might appear in email written by people less familiar with him, who wrote his difficult-to-spell surname as “Robert Shakleton” or “Robert Shackeltin.” If Bobby was the topic of conversation in emails written by the company’s office in Tokyo, he might be referred to as “ロバート・シャケルトン” (Robert Shackelton).

In this case, entity search might finds matches (e.g., “Robert Shakleton” vs. “Bobby Shackelton” vs. ロバート・シャケルトン) that a human sifting through documentation might not think was a match.

One of the keys to this use-case is identifying the entities in the unstructured information, and linking them against structured information. By providing name-search functionality on top of entity extraction, a lawyer or investigator can quickly and efficiently find variant spellings that actually exist in the document set.
Below illustrates how these technologies could be deployed in an e-discovery system, with the name entered in the upper left and the system searching for possible name matches.

5. FINANCIAL COMPLIANCE

Financial institutions must comply with government regulations and limit reputational risk by implementing “know your customer” programs and screening customers against watch lists. They also need to recognize customers that may present higher-than-average risk to the organization. Entity search, using a high-quality name matcher, can be used to match customer names against watch lists, and politically exposed persons (PEP) lists. As part of due diligence, these financial organizations must also survey adverse media—news that may contain risk-relevant information—for each of their customers. Combining entity extraction with name matching enables these organizations to quickly identify when a customer is mentioned in adverse media to help inform the risk level a particular customer presents.

At a basic level, name matching can be applied to structured sources, like checking the “name” field of a new customer application against a list of sanctioned or high-risk individuals.

At a more advanced level, screening current and potential customers for adverse media coverage requires searching news and media feeds to discover when customers are mentioned.
Although a keyword search through media may locate “Ted Kaczynski,” it may not find “Theodore Kaczynski,” “Ted Kazinsky,” or “Teddy Kazynsky.” However, by combining entity extraction (to locate names in the adverse media articles) with name matching (to match customer names against those found in adverse media), organizations can quickly identify when their customers are mentioned, even if their names are spelled differently.

In this risk-sensitive market, applying entity extraction and name matching will result in a more comprehensive solution, lowering risk for financial institutions.

6. SOCIAL MEDIA MONITORING

Social media monitoring applications are designed to measure buzz and track conversations about brands and products. They also provide a way for organizations to identify opportunities to engage with their customers via social media, and track their success with marketing campaigns.

In the case of brand monitoring, a brand name such as “Krispy Kreme” might also be spelled “Crispy Cream” or “Krispy Cream” in social media postings. To effectively monitor posts about this brand, the monitoring system must recognize that these refer to the same thing. A powerful entity extractor can be used to identify each of these as entities in social conversations, and feed that information to a name matching engine.

The name matching engine would connect those mentions as being similar spellings, allowing everything from a person’s personal recipe for “Krispy Cream” doughnuts to be found alongside discussions about Krispy Kreme doughnuts on FaceBook, Twitter, or Yelp.com.

Customer engagement is an equally important area. Companies want to be able to connect to customers who may be asking questions or complaining about a company service or product via social media.
In either case, entity extraction will locate mentions of the specific brand, product, or company in social media posts, and name matching will find all similar spellings. This means that when a user searches for “Krispy Kreme,” the name matching engine would return “Crispy Cream,” “Krispy Cream,” or “クリスピー・クリーム.”

7. GOVERNMENT INTELLIGENCE & LAW ENFORCEMENT

Both intelligence analysts and law enforcement officers deal with a variety of information, ranging from highly structured—a database of visa applications—to unstructured field reports, message traffic, Internet webpages, chat room conversations, and more. Navigating any one of these complex data sources is challenging, but manually identifying links between data sources is nearly impossible without the right technology.

In the famous case of the Christmas bomber in 2009, there was plenty of finger-pointing as to why Umar Farouk Abdulmutallab was not put on the no-fly list given the information about him available to U.S. intelligence agencies prior to the incident. The investigation into the near-bombing concluded that Mr. Abdulmutallab might have come up for consideration for the no-fly list had all the evidence about him been connected. However data such as his U.S. visa application, his listing in the Terrorist Identities Datamart Environment (TIDE), his father’s warning about his son’s potential radicalization, and other information did not come together, in part because of variations in how his name was spelled in each data source.

In this scenario, entity search across these various databases could have provided analysts the ability to link these various pieces of information—with entity extraction identifying his name in unstructured diplomatic cables, and name matching technology linking the entities against the other spellings of Mr. Abdulmutallab’s name in the structured databases, presenting a more complete picture than was available at that time.

8. ENTERPRISE-READY TECHNOLOGY

Basis Technology offers an enterprise-level suite of text analytics components in the Rosette® linguistics platform. The capabilities described in this document are provided by the Rosette Entity Extractor (REX) and the Rosette Name Indexer (RNI). The entity extractor combines several approaches to ensure high accuracy and performance for 16 different entity types in more than 15 languages. The name indexer—combining name search and name matching—uses an integrated suite of algorithms that handle a wide range of name matching problems, while maintaining high throughput. Besides spelling and phonetic variations, Rosette also matches names across languages and common variations including typos, nicknames, missing spaces, missing name components, and more.

Rosette can serve as the basis of new applications or integrate quickly into existing systems to provide advanced capabilities that save you and your customers time and money.

9. EXPLORE FURTHER

For more information or to request an evaluation, please call us at 617-386-2090 or 800-697-2062, or write to info@basistech.com. We will be happy to assist you in evaluating the performance of our products on your data.