TEXT CLASSIFICATION ADVANTAGE VIA TAXONOMY

A white paper by Seth Grimes, Alta Plana Corporation

I. INTRODUCTION

This paper addresses the numerous challenges of automated text data handling, for applications that range from social intelligence and user-experience design to search marketing to and contextual advertising. It explains advantages delivered by use of a particular technology, taxonomy, for a key technical step, classification.

This paper describes the technology and application scenarios: The *What* of taxonomy and alternative approaches, and *How* and *When* to apply taxonomy-centered solutions for optimal results. The *Why* of taxonomy is the business advantage that stems from accurate and reliable automated processes: Customer satisfaction, sales conversion, efficient support, retention, and monetization.

II. BUSINESS AND TECHNICAL CHALLENGES

We live in a world of high-velocity, high-volume, online, social, enterprise, and device data. The *Internet of Things* (IoT) is an emerging reality. Smartphones, sensors, machinery, and servers generate extraordinary volumes of data. The IoT is complemented by an *Internet of People* (IoP) – consumers, influencers, our communities, competitors, and the broad public – you and me. We create and consume a growing amount of diverse content, across multiple devices and often while on the move, whether for business or personal purposes. We share our news, views, and needs, via messaging and apps, review sites, and online and social media as well as via in-person encounters and traditional documents and channels.

Understanding consumers, personalizing experiences

It is common practice to track and measure customers' and public activities online, on-social, and in the public sphere, subject to privacy constraints, and when appropriate, to engage, to respond.

The aim of public-facing organizations, whether commercial or non-profit, is to use the data generated to understand individuals' needs and better serve them. Business-to-business functions face similar needs. Automation is a necessity; nonetheless, our audiences expect personalized, one-to-one experiences. To support appropriate, efficient, and productive user, customer, and business interactions, accurate, reliable, and flexible methods are essential.

Diverse technical challenges

There are diverse technical challenges. We can summarize them in a few bullet points. When we automate text processing, we must:

- Accurately identify content – articles, ads, product descriptions, service information, reviews, and the like – relevant to given audiences, outlets, individuals, and occasions.
- Deal with the vagaries of natural language – the misspellings, fractured grammar, slang, and idiom that are common in everyday written and spoken language.

- Match content analysis and delivery to *situational* needs, accounting for context, profile, demographics, and behaviors.
- Move beyond keywords, dealing with ambiguity and taking into account interest categories, detailed attributes, associated topics, and related concepts.
- Operate in *right-time*, whether the need is *low-latency* processing of streaming data, fast response to an online inquiry, or trend detection and forecasting via data mining and predictive modeling.

Technology choice

An array of technologies attempt to help you meet these business and technical challenges. Some have been proven through years of experience. Others are experimental. **It is critical to choose the right method, sensitive to the nature of your data and analytical or operational needs, with the reliability and performance you require.**

III. TECHNICAL APPROACHES

Technical challenges may seem daunting, answerable only by complex systems of language rules, statistical algorithms, or highly sophisticated machine learning. Yet in many instances, straightforward, simpler techniques perform best. Taxonomy is a textbook example.

Let's look at taxonomy and at other language-analysis methods, before taking on application scenarios. We'll start the look at taxonomy with a definition.

Taxonomy 101

Taxonomies group things of particular types or categories into hierarchies. They model relationships – "this is an instance or example of that" – often to several levels of depths, from detailed to general. Taxonomy's origin dates back over two thousand years, to Aristotle and other ancients. They sought to describe, via categories, everything that exists or can exist. Carl Linnaeus's 18th-century classification of the natural world is a wonderful, well-known example. Each branching of the tree distinguishes subgroups that, while they share enough characteristics to be grouped together at the branch point, may be differentiated by other, more-detailed characteristics. Take the example of mammals, distinguished from birds, insects, fish, and other animal classes by milk production.

How would an automated process apply taxonomy? By matching words and terms found in a given text item – in a Tweet, e-mail messages, article, or document – to a taxonomy, preferably accounting for usage density. The aim is not only to classify the item, but also to boost item usefulness – for search, classification, and information usability – by identifying (per the fourth technical challenge, above), broader categories, detailed attributes and related concepts.

Taxonomy + technology

The capacity and performance of modern computing allows for the creation of deep and broad taxonomies that capture the objects and types within a given industry or business along with their many attributes. The same Web-crawling technology that harvests new content for Google indexing – and the same social media monitoring technology that underlies listening and engagement solutions – can be applied to identify new, topical *things* for inclusion in industry-domain and business-function taxonomies. Modern technology

allows taxonomy-based solutions to be applied for a diverse set of challenges that answer just the sort of online and social business needs we have examined.

Contrast: Linguistic analysis, statistics, and machine learning

The information content of documents, messages, speech, and search queries (to name a few of the many forms text takes) is communicated via words, organized into phrases, sentences, narrative, and conversations. Let's put aside emoticons and emoji and creative punctuation (!!!), while word morphology (tense, conjugation, declension, plurals, and gender agreement), misspellings and abbreviations, and every sort of grammar and syntax used "in the wild" remain in-bounds.

Parsing and counting

Some *natural language processing* (NLP) methods seek to decode subject, object, and verb – this approach is called *part-of-speech tagging* – in order to understand relationships among the things the words refer to. A more basic first step, however, after *tokenizing* individual words, is *named entity recognition* via look-up of person, place, company, and product names in lexicons or gazetteers. Some tools rely on language rules that encode generalized associations among words; for instance, "Mrs. <word>" probably indicates a person while "state of <word>" may name a place (or a condition such as confusion). Rules trade the exactness of list look-up for flexibility that may identify things not already on your list.

More sophisticated is *theme or topic extraction* via simple word and term counts or via statistical clustering. We might use word adjacency or co-occurrence to infer attributes, and further, *lexical chains* and *word networks* help us decide the contextual meaning of ambiguous terms. Other computational linguistics algorithms resolve *coreference*, multiple ways of referring to a given thing (e.g., Barack Obama, Mr. Obama, the President, and in certain cases, the pronouns "he" and "his.")

Decoding language is hard; decades of research and development have been dedicated to the task. As a result, natural language processing approaches may be quite involved, especially when they need to deal not only with multiple human languages, but also with bad grammar, misspellings, acronyms, slang, and sarcasm. As an automation end run, or to facilitate creation of linguistics lexicons, networks, and rules, we have machine learning.

Machine learning

The term *machine learning* covers many methods. A basic distinction is *supervised* methods, which build models from *training data*, and *unsupervised* methods, where the software creates a classification model from whatever the algorithm determines is statistically interesting. **Machine learning can work quite well, if you have enough model-building data and also if you retrain your models to keep up with new terminology**. **But machine learning on its own will never deliver the descriptive precision – the exact modeling of a knowledge domain – possible with a well-crafted taxonomy.**

IV. SMART SELECTION

Recapping, our analytics aim is to fully understand customers', prospects', and market needs in order to optimize product and service delivery, creating excellent experiences and boosting satisfaction, loyalty, and profitability. Our audiences expect personalized, one-to-one experiences.

Every aspect of the customer journey – and of supporting market research, product and service design, and demand forecasting – relies on accurate classification technology. Classification is at the heart of describing, modeling, and predicting individuals' interests, behaviors, and affinities, based on pattern detection and demographic and behavioral profiling. It enables software to create topical "semantic signatures" of online and social content, in order to facilitate automated matching and recommendations, and also to avoid negative associations.

Use Cases

There are many situations where taxonomy-based classification is the most accurate and effective classification approach. Top-value scenarios include:

- Online Advertising: Match the taxonomy-derived *semantic signature* of a Web page or social content to a precisely tagged ad, and you have precision ad placement. **You can rely on exactness not possible via statistical topic extraction**, which won't model the important conceptual relationships captured in a taxonomy. And if you don't want a car ad matched to an article on an accident? A taxonomy can record negative associations that support "don't show" instructions.
- Recommendation Engines: eBay's "See what other people are watching" and Amazon's "Customers who viewed this also viewed" are based on behavioral patterns. Useful, but what if a product is new or less popular (that is, there's little viewing history) or your online storefront hasn't attracted the Amazon-scale visit volume needed to build useful behavioral models? **Taxonomy based classification doesn't rely on behavior modeling**. It can associate, for example, screen protectors, cases, earphones, and other accessories with cell phones, by manufacturers, model, and even color, with unequaled precision.
- Social Media: Social practices have focused largely on listening and engagement, with predictive analytics – an essential tool for next-generation online commerce, market research, brand and product management, and even government policy – only infrequently deployed in the social sphere. **A deep and broad taxonomy would classify and associate social users' interests, behavioral patterns, and intent, advancing the cause of social in business.**
- Customer Service/Support: Imagine a contact-center agent interaction, or an e-mail message requesting product support. We can speed response to these and similar inquiries if we can automate routing of the inquiry and identification of content that answers the need and also associated components and potential issues. These functions call out for a mechanism that interrelates items (products and services and their components), attributes (size, features), concepts (repair, refund, failure), and content (guides and instructions). **Taxonomy provides that mechanism, via integration into customer-interaction systems and search engines.**
- Consumer Insights: Consumers use a very rich vocabulary of nicknames, terms, and slang in referring to products and services. **Via a taxonomy, you can map "in the wild" language used on online and social platforms and in e-mail and surveys to a *controlled vocabulary* of terms, in order to normalize the data you've collected, to support analysis and insight extraction.** As an aside: Location is an important

consumer-insight component. Taxonomies provide an excellent mechanism for handling geographic reference data.

- Demand Forecasting: A deep taxonomy may capture both high-level information about brands and product categories, and detailed specs covering features and attributes. **A taxonomy's hierarchical structure allows for classification of forward-looking market insights at multiple levels.** At a category level, we might ask "What class of toys will be hot next Christmas, given advance buzz?" At a feature level, the question becomes "How are consumers' mobile phone battery-life expectations trending?" You can get multiple roll-ups from a single set of information sources without the expense of complex rule sets or of training and managing multiple machine-learning-built models.

- Competitive Intelligence: A statistically rooted method, fed enough data, may infer that iOS and Android are both mobile-device operating system. Through *co-occurrence analysis*, it may determine which vendors sell Android devices. A combination of *named-entity recognition* and *measure-value extraction* will pull sales figures by device or vendor out of published news articles: impressive but insufficient if your need is to aggregate information that responds to multiple attributes, per this example. **By contrast, a taxonomy may allow you to short-cut the statistical modeling – human experts are typically great judges of topical information-organization** – while supporting roll-up to comparable classes and categories at the level of detail you desire, an important supporting capability for competitive intelligence.

- Search & Site Retargeting: Have you searched online for a product, or visited a vendor's Web site, and then been repeatedly subjected to product ads when visiting other sites? That's search and site retargeting at work, a special form of advertising delivery. You'll see retargeting ads even if you've already made a purchase – even if you don't search again for that product or visit product-relevant pages – because the tracking networks are ignorant of context. Further, tracking is a challenge on non-cookie based smartphones so that accurate contextual mapping of content is particularly vital. **A taxonomy-based solution will allow targeting based on the *semantics* of the content and not based only on past behaviors.**

These scenarios include direct application of the technology, through domain-adapted user interfaces and also as part of larger analytically reliant solutions, where analyses are invoked *as-a-service*, via *application programming interfaces*, with workflow managed by a purpose-built solution. Keep in mind that implementations are not either/or. **A well-constructed taxonomy will complement other methods, handling classification needs when other methods fall short**.

V. WHAT TO LOOK FOR IN A SOLUTION

Different tools have different capabilities, different uses and strengths. In application of taxonomy (or any other technology) to text analysis tasks, you'll want to identity the solution that best meets your needs. Here are attributes that will influence your selection:

- Domain suitability: A system designed for medical research, covering pharmaceuticals, diseases, clinical symptoms, and anatomy, would be an odd choice

for hospitality-industry market research or customer service. Adaptation may involve interfaces and algorithms.

- Task suitability: We accomplish tasks via multiple steps, often involving a series of decisions. A graphical user interface will offer ease-of-use in accomplishing the tasks it was designed for, but it may dictate a certain workflow and output choices. So for certain tasks, you'll want the flexibility of building a solution from a componentized, or as-a-service, technology choice.
- Scope: Even a suitable choice may not be a best choice. For instance, a taxonomy that captures hospitality terminology will fall short in analysis of travel reviews if it lacks food service and restaurant coverage.
- Precision: Detail enables exact classification, the ability to differentiate based on fine-grained characteristics. Look for a level of precise that provides complete domain coverage (breadth) and enumerates all variations and attributes (depth).
- Accuracy: Simply put, is the software or taxonomy publisher's work correct?
- Currency: Is the taxonomy or model frequently refreshed with new categories, nodes (whether companies, products, or people), and attributes?
- Ease of Implementation: Does your method of choice work 'out of the box,' across all subjects – a distinctive advantage – or is model training (e.g., for machine learning) or rule-writing (language-engineering approaches) required**?**

These attributes reflect familiar concepts. Domain and task suitability are related to relevance. What we call scope here is similar to search *recall*, or result-set completeness. Accuracy and currency concerns are universal and independent of the method.

There is a temptation to choose a solution based on source or on endorsements. While these are important factors, also consider the provider's industry experience, objectivity, and record providing exemplary customer service.

VI. ABOUT THE AUTHOR

Seth Grimes consults on business applications of text analytics, sentiment analysis, and data visualization. He founded Alta Plana Corporation in 1997 and the Sentiment Analysis Symposium conference series in 2010. Follow him on Twitter at *@SethGrimes*.

VII. ECONTEXT

This paper was sponsored by **eContext**, a SaaS text classifying technology. eContext discovers insights and intent in vast amounts of data to give brands, publishers and marketers unique and valuable advantages. eContext is a division within Info.com

