Opening Wikipedia's Window to our Collective Mindscape
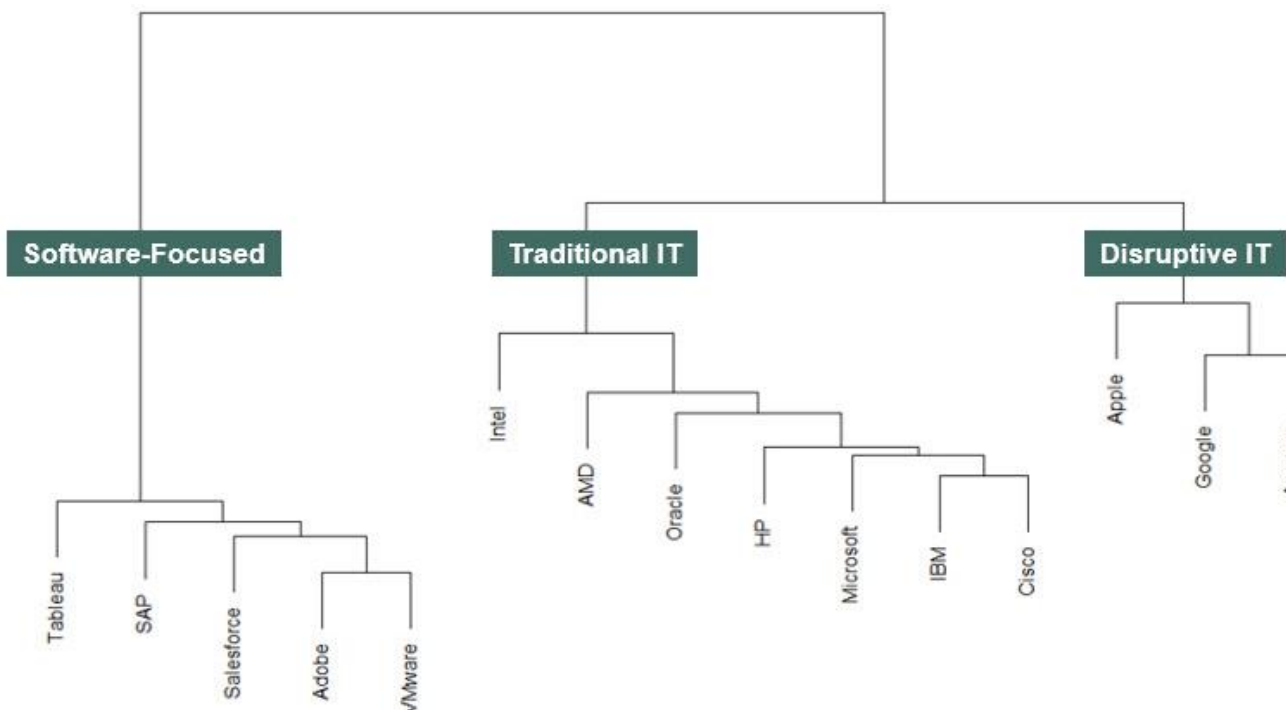
By Michael Kelly, Ph.D.

With more than 5 million articles and 6 million daily users on its English site alone, Wikipedia is a key resource of the globally connected age. Beyond that, its vast scope and crowdsourced development also means that it's a potential window into our "collective mind."

This collective perspective is difficult to see from our everyday encounters with Wikipedia, in which we delve into individual articles one-at-a-time. With appropriate tools, though, we can rise above and aggregate across specific articles for a "20,000 foot" view.

Text analytics and statistical clustering algorithms are examples of tools that can group Wikipedia articles or other text into categories based on overlapping vocabulary. The output can be quite revealing about how our culture implicitly organizes a category – and how those organizing principles might change over time.

As an illustration, we applied these tools to Wikipedia entries for companies that are broadly in the information technology space, such as Apple, Google, IBM, Microsoft, and SAP. These tools will group companies close together to the extent that their Wikipedia entries share words, and these shared words appear at similar rates. In the resulting tree diagram, we see that three clusters of companies stand-out, which we've labeled "Software-Focused," "Traditional IT," and "Disruptive IT" based on market orientation. (While these tools deliver potential segments, it's up to us to interpret them.)

There are no "silver bullet" words that are responsible for these segments. Instead, a number of nuanced differences combine into higher-level themes to grow the tree, such as those in the table below. While some of these patterns may be expected, others are more subtle indicators of differences in company style or focus.

Themes that Appear More Often in Wikipedia Articles by Segment	
Traditional IT	• Traditional computing (e.g., hardware, computing, computer, circuit, silicon, semiconductor, technology) • Corporate terms (e.g., CEO, chairman, leadership, corporate, companies, president, market(s), business, firms) • Formal business verbs (e.g., filed, release(d), develop(ed), announced, agreed, named, completed) • Law/regulation (e.g., government, court, legal, rights, patent, case)
Disruptive IT	• Media (e.g., CNN, Bloomberg, WSJ, Forbes, Guardian) • Internet age terms (e.g., online, website, blog, smartphone, startup) • Success terms (e.g., top, main, biggest, largest, increase, original) • Short "active" verbs (e.g., create, launch, make, talk, take, sell) • Focused time references (e.g., specific months of the year, quarter)
Software-Focused	• References to competitors (e.g., companies in tree plus others) • Software terms (e.g., database, Java, app, application, CRM) • Cloud-related terms (e.g., cloud, virtual, virtualization) • End-user devices (e.g., iPhone, iPad, Macintosh, Kindle) • Processors and latest storage technology (e.g., Opteron, Pentium, Athlon, GPU, flash, SSD, solid-state)

In a business context, mindscape analyses like these can help companies understand at a high level where they are perceived to fit in the cultural market space, and the specific themes associated with their "neighborhood." Business leaders can then ask if the "company they keep" – or are perceived to keep – aligns with their self-image and objectives, and whether they should try to address a significant misalignment.

About the Author

Mike is a Senior Group Director at NAXION who designs and manages major engagements for clients seeking to develop B2B and B2C business strategies based on customer insight and advanced market analytics. Leveraging his skill in devising new modeling techniques and integrating multiple data streams, Mike has helped clients in Information Technology, Energy, Consumer Electronics, and Manufacturing to optimize pricing strategies, guide product bundling, sharpen targeting, and prioritize service improvements that drive customer loyalty. Mike has notable subject matter expertise in the fields of cognition and computational linguistics, where his highly regarded academic work has been a platform for innovation on behalf of clients, while building the firm's intellectual capital in advanced methodologies.

Mike Kelly, Ph.D., Senior Group Director
mkelly@naxionthinking.com 215.496.6842