Data-mining Wikipedia - a new frontier of insights

| By Michael Lieberman

snapshot

Michael Lieberman explores how network analysis can help researchers understand brand, business and marketing relationships.

The definition of Wikipedia is "a free, web-based encyclopedia project which contains information on a wide variety of subjects which are added by contributors from all over the Internet." The world's eighth-most trafficked public website, Wikipedia is an open-source wiki software and, as with all wikis, it allows everyday users to create and edit webpage content in any browser.

Business and brand use of Wikipedia is ubiquitous. Every business/brand/city/enterprise has a Wikipedia page. Each wiki page is essentially comprised of a concept and a relationship to other concepts.

And now, thanks to network analysis, these pages can be mined by insights professionals to describe brand spaces, understand the conversation around an advertising campaign or explore the social landscape of a product.

Network analysis software allows us to feed it a Wikipedia seed page – the seed page is the first page the software looks for, like "Marketing" or "IBM." The next step is to define other wiki pages that are mentioned or connected to the seed. Once established, the software determines where the other pages are connected to each other. In network parlance, this is called a neighbor-neighbor network.

Once the links are established, a macro that clusters connections is run. This categorization of network links is not different than, say, a factor analysis, which looks for underlying structures of data using statistical algorithms. The software then visualizes their positions and creates links for each wiki in the map and where they stand vis-à-vis the seed page. This creates an ecosystem of relationships with immense research value.

In this article we will provide a brief background to the Wikipedia network space. From there we will run through the process of an insights-mining exercise (for example, the connected articles of an industry) using the knowledge networks produced by successive Wikipedia maps, followed by an explanation of next steps to give the shape of the social conversation in-depth context.

Define it

Strategists often talk about wanting to control the social media conversation around a brand or company. Given the dispersion of the landscape of social media, the first challenge to control the conversation is to define it. One method is to explore the shape, or frame, of its connections. This is accomplished by the above-described knowledge network.

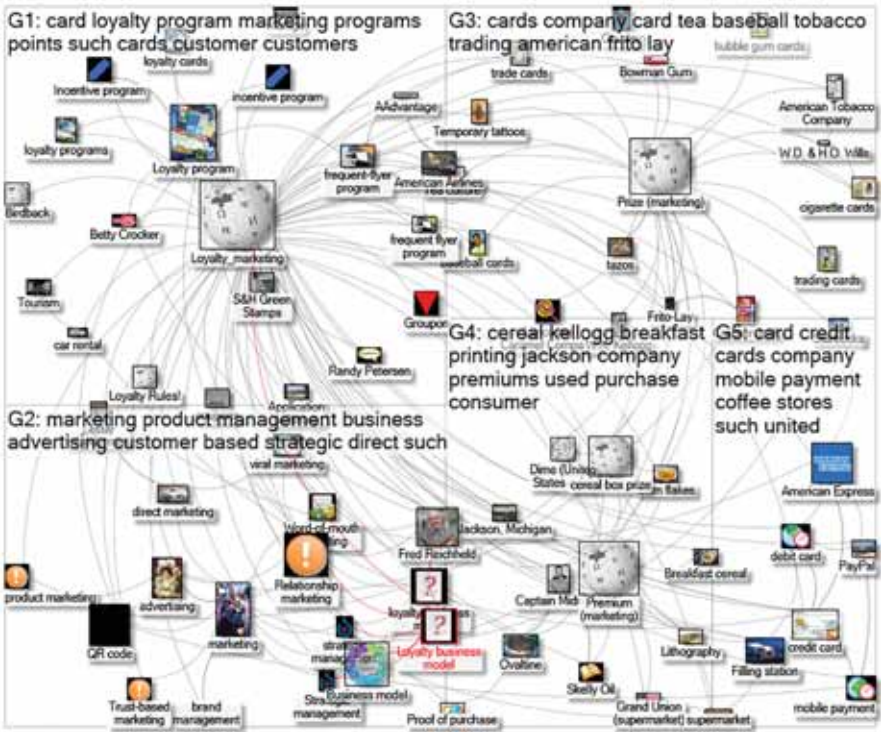
A knowledge network is a group of people who are equipped with tools that enable collaboration, interaction and sharing of work-related experience, know-how, expertise and resources. Every knowledge network is created under a unique set of circumstances.

Marc Smith, founder of NodeXL, a network visualization tool that includes access to social media network data importers, advanced network metrics and automation, says, "A Wiki-

www.quirks.com/articles/2021/20210104.aspx

Figure 1

Figure 2



pedia network map reveals webs of interrelated concepts and entities that are not necessarily obvious. And clusters that are not obvious. These maps can be semantic groups, ontologies [entities that really or fundamentally exist for a particular domain of discourse] or knowledge graphs. Looking at a wiki page is like looking at a leaf – not the branch, not the tree, not the forest. Network analysis allows us to see the relationships between a page and its neighbors and to see clusters of relationships among topics."

In other words, when given a seed article, what other articles or subjects are connected to that wiki? It's the connections that tell the story and those connections can be represented by a series of Wikipedia network maps.

As an example, let's look at Figure 1, which shows a map using the seed page on Wikipedia for "Marketing_Research." The size of the sphere indicates its centrality, that is, how many other wikis are connected to it. The seed article, "Marketing_Research" has the largest sphere. The spheres are also clusters according to how they relate to each other. Thus, from this map we can learn:

G1: Highlights the foundations of the marketing research industry; which other Wikipedia pages are directly connected to the broader seed page and vital uses of marketing research such as consumer behavior and social network potential.

G2: Focuses on the interconnections of various analyses such as pricing, brand marketing and brand awareness.

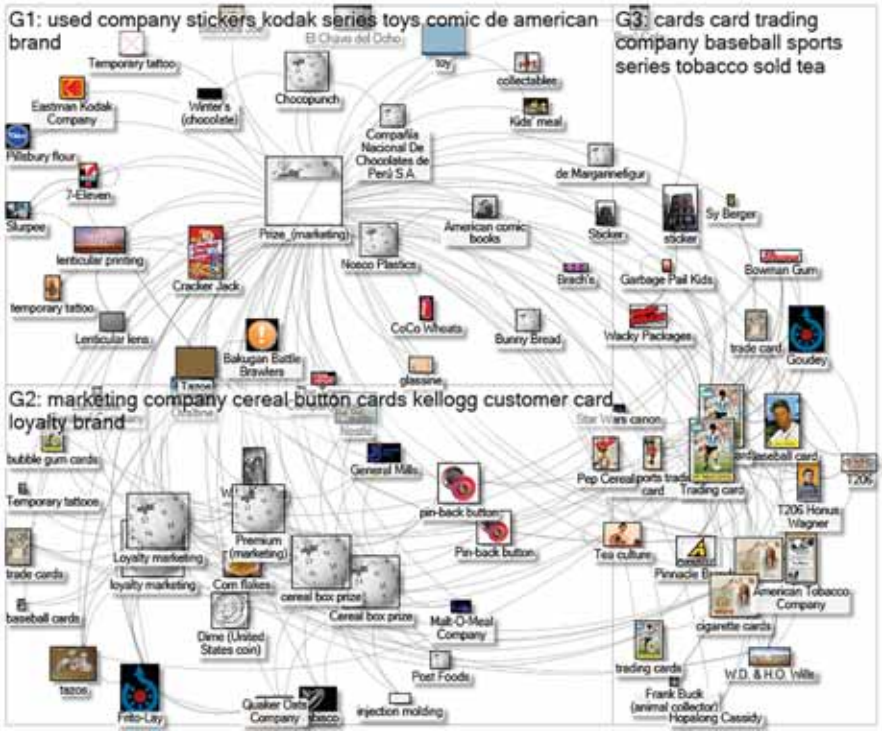
G3: A third group we can call the "quantitative group," as it shows that many Wikipedia pages that are connected to the seed page are connected to each other, such as new product development, price elasticity and conjoint analysis.

G4: Highlights the advertising research connections within the marketing research knowledge network.

Loyalty marketing, a connected part of the marketing research knowledge network map, is another way to illustrate the research process of media wiki mining. Branding, product marketing and loyalty marketing all form part of the customer proposition – the subjective assessment by the customer of whether to purchase a brand based on the integrated combination of the value they receive from each of these marketing disciplines.

The research process for a Wiki-

Figure 3



pedia project is a sequential series of network analyses. As stated before, it begins with an industry snapshot. We would then attempt to drill down and examine comparative relationships within maps. In our example study, let's say we are trying to create an in-depth knowledge space for loyalty marketing efforts. The natural next media wiki map would be of "Loyalty Marketing," which is shown in Figure 2.

The interwoven network loyalty marketing media wiki pages begins with the first group:

G1: Linking to various types of loyalty marketing and who is undertaking them.

G2: The second group of linked Wikipedia pages (G2) explores relationship marketing by means of direct or viral marketing, word-of-mouth and trust-based methods.

G3: The third group (G3) features connected wikis of prize marketing; if, for example, a brand did not realize the prize marketing is connected to loyalty marketing.

G4: A subgroup of breakfast foods and premium marketing programs.

G5: Credit card company loyalty programs are linked on Wikipedia.

The continuation of the study would be perhaps to create a series of maps around specific companies, comparing their individual knowledge spheres. For example, setting the seed for "Prize Marketing" would result in the map in Figure 3.

With comparison of multiple maps we can then form insights into the section of the report. It would be difficult to display 10 graphics in this article. However, here is an outline of the course of a given study:

- industry media wiki knowledge network maps;
- client and major competitor media wiki maps;
- current comparative social conversations around the industry, client and competitors;
- which companies are involved with prize-marketing strategies.

A deeper dive into this research project is beyond the scope of this article but one must not forget that each Wikipedia page contains information and details of its elements. Furthermore, the social conversation around a topic could be followed on other social media platforms.

Vital aspect

Media wiki networks have been around since the turn of the century and they are still a vital aspect of the online behavior of commercial entities, private individuals and governments. We expect the availability of tools such as NodeXL to have a positive impact on marketing research in a previously untapped space.

As mentioned above, dynamic media wiki analysis is a fruitful area of study, as is research into approaches for jointly analyzing media wiki and text-content data. Companies have learned to harness the power of thought leaders, experts and influencers to promote their products and they can use media wiki web space visualizations to better understand the shape of those networks.

Michael Lieberman is founder and president of Multivariate Solutions, a New York consulting firm. He can be reached at michael@mvsolution.com.