What do Universities, Congressmen, and Baseball Teams have in Common?

Raul E. Valdes-Perez, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213

Well, this writer's city of residence has at least one of each. But my aim is to point out that in each case, it is interesting to know how any single university, member of Congress, or baseball team is different from all other universities, members, and teams. That is, finding a niche for any of these individuals is interesting because it responds to a basic human desire to make sense of the world in terms of individual idiosyncrasies, expressed concisely. This is true of politics, sports, and education, and is no less true of business, science, and other professions.

For a long time I have observed the tendency, perhaps in our American popular culture especially, to make niche statements. For example, upon arriving in Pittsburgh in 1986, I learned that the University of Pittsburgh's famous Cathedral of Learning is the tallest academic building in the world outside of Russia (Moscow State's is the tallest). Driving back from New England this past summer, I came across a road sign advising that I was passing the highest point on Route 80 East of the Mississippi. On its web pages at www.saic.com, Science Applications International announces itself as "the largest employee-owned research and engineering firm in the nation." These niches are all somewhat noteworthy, although we have all seen abusive examples, such as a community that puts up a sculpture and, rather than commenting on its beauty or significance, advertises that it is the largest bronze statue made by a left-handed sculptor in this century, or some such nonsense. Finally, in October of '98 I was perusing the web pages of the University of Miami in preparation for a trip to South Florida, and this exquisitely crafted niche caught my eye: "The University of Miami is the youngest of 23 private research universities in the country that operate both law and medical schools." The impact of this statement, combined with a long-standing interest in niches generally, and mixed with insights from some computer science research on problems in anthropology (making sense of the kinship terminology found in different cultures) led to the sudden realization that, given the right data, such niches could be generated effortlessly with the aid of computers. So, given data on university characteristics, congressional voting records, and baseball statistics, one can use a computer to concisely articulate what is unique about individuals in any of these areas.

An important issue is what forms of niches are acceptable, and we can turn to the considerable body of human practice to answer this. A clue is present in the four niches listed above. In each of them, one numerical attribute is mentioned: tallest, highest, largest, and youngest. The remaining attributes establish a context within which the numerical attribute is made extreme. Thus, Pittsburgh's building is not the tallest in the world, but only among academic buildings excluding Russia. Miami is not the youngest university overall, but only within the stated context, and so on. So we see that niches having one numerical attribute are common and understandable. Actually, if they weren't understandable, they wouldn't be common. We still need to consider whether niches having no numerical attributes, or those having two or more, are acceptable.

A niche that has only context-establishing attributes is clearly acceptable, since it will be of the form "is the only" and this form is ubiquitous. For example, the Timken Museum of Art advises on the web that "Rembrandt's Saint Bartholemew is the only painting by that Dutch artist on display at any museum in San Diego." The Open Zoo in Singapore is the "only zoo in the world which offers a regular 'Breakfast/Tea with an Orang Utan' programme for guests." Not only English is implicated: a web guide to France's Lyon claims that the city has "le seul zoo gratuit d'Europe."

The subtler challenge is deciding whether niches should be allowed to have more than one numeric attribute. We can distinguish two cases that are illustrated with these phrases "is larger and older than ..." and "is either larger or older than ..." The first of these is really a simple combination of two niches, so we can disregard them and consider finding these a separate challenge. The second is not a mere combination of two niches, but only a logician could understand it (which is itself a niche for logicians). So we reject them as too complicated. Concluding, niches are acceptable only if they mention at most one numeric attribute. (One notices, though, that some context attributes are derived from numerics, but they don't count against the ceiling of one. For example, it would be acceptable - even if uninteresting - to claim a film as the

longest-running top-ten grossing film of all time. "Top-ten grossing" determines the context.)

Having decided what niches are OK, we ask what makes one niche better than another. In most cases, conciseness is the key: making use of the fewest attributes possible. It is generally uninteresting to say that some individual is the only one that possesses twenty different attributes. We further ask what makes one niche better than another that is equally concise. Now the answer rather depends on the specifics of the area, and it's hard to find some general principle. Conciseness is the key goal, although not the only one.

Are niches really that widespread? One can use that phenomenal resource - the web - to find out. An Altavista search on the phrase "is the largest company" turns up 854 web pages, while "is the smallest company" still matches 27. Other examples are: "is the only museum" (1024 pages), "is the oldest university" (710), "is the largest sculpture" (19), "is the winningest" (1359), and so on. Again English is not alone: "es la mayor empresa" (Spanish for the largest company) hits 99 pages.

I have written a computer program called PickNiche that finds niches using the above principles, given data on individuals and their attributes. (As far as I can tell, it is the only program in the world that does this.) It compares a selected individual against every other, calculates their differences, and then minimizes the expression of these differences in order to find concise statements.[1] The final output is a grammatical English sentence. I ran the program on baseball teams, congressional voting records, and U.S. universities, using data taken almost entirely off the web. Some examples follow; larger collections are on my web page at http://www.cs.cmu.edu/~sci-disc/pickniche.html.

Baseball. What's special about any of the 2,297 teams that have played major league baseball since 1871 through 1998? One learns, for example, that the memorably bad but fun 1965 New York Mets had the most losses (112) of 897 teams that had a million in attendance. Or that last year's San Diego Padres struck out the most (during the regular season) of any team that lost in the World Series. The 1998 New York Yankees had the most wins (114) of 1,914 teams that had a team ERA above 3.00. The 1975 champion Cincinnati Reds had the most wins (108) of 1,313 teams that did not have a 20-game winner and did not have a pitcher with 200 strikeouts.

Universities. What's unique about each of the 236 major U.S. universities? Some interesting - and difficult to think of - niches are that MIT has the most members of the Institute of Medicine (an honorary society like the National Academy of Sciences, but for Medicine) of any major university that doesn't have a medical school. The University of Illinois (Urbana-Champaign) graduated the most Nobel Laureates (8 as of early 1999) of 92 major non-California universities that play Division IA football. Wisconsin, Madison graduated the most (13) major-company CEOs (Forbes 1994 data on 800 CEOs) of any top-ten party school, Notre Dame the most (10) of the 15 major Catholic universities, and Southern Methodist the most (10) out of 110 major non-research universities outside the Ivy League.

Congress. How is one Congressman different from all the other Congress members in terms of their voting records? I downloaded voting data on the 1999 session of the House of Representatives and obtained niches such as: among the 97 Democratic Members of Congress who voted for the Bankruptcy Reform Act, Congressman Gonzalez was the only one who voted against the Student Results Act. Of the 134 Members of Congress who voted for the Civil Asset Forfeiture Reform Act and against both the Regulatory Right-to-Know Act and the Workplace Preservation Act, Congressman Boehlert was the only Republican among them. All of the 436 Congress members in the dataset had their own niches except four.

Human beings naturally try to make sense of their world by seeking out what is distinctive about individuals, events, institutions, and so on. Statistical theory and data mining have their origin in scientific methods and traditions, which test hypotheses, look for causes, propose models, and classify or profile groups; mere facts are grist for these mills. Facts are to be observed or measured, but are usually not themselves targets of complicated reasoning or discovery. Maybe this is why niche finding has been overlooked for so long.

[1] The methods are patent pending.