Making search better in Catalonia, Estonia, and everywhere else

Tuesday, March 25, 2008 at 4:02 PM ET

Posted by Paul Haahr and Steve Baker, Software Engineers, Search Quality

(Cross-posted from the Official Google Blog)

We recently began a series of posts on how we harness the power of data. Earlier we told you how data has been critical to the advancement of search; about using data to make our products safe and to prevent fraud; this post is the newest in the series. -Ed.

One of the most important uses of data at Google is building language models. By analyzing how people use language, we build models that enable us to interpret searches better, offer spelling corrections, understand when alternative forms of words are needed, offer language translation, and even suggest when searching in another language is appropriate.

One place we use these models is to find alternatives for words used in searches. For example, for both English and French users, "GM" often means the company "General Motors," but our language model understands that in French searches like seconde GM, it means "Guerre Mondiale" (World War), whereas in STI GM it means "Génie Mécanique" (Mechanical Engineering). Another meaning in English is "genetically modified," which our language model understands in GM corn. We've learned this based on the documents we've seen on the web and by observing that users will use both "genetically modified" and "GM" in the same set of searches.

We use similar techniques in all languages. For example, if a Catalan user searches for resultat elecció barris BCN (searching for the result of a neighborhood election in Barcelona), Google will also find pages that use the words "resultats" or "eleccions" or that talk about "Barcelona" instead of "BCN." And our language models also tell us that the Estonian user looking for Tartu juuksur, a barber in Tartu, might also be interested in a "juuksurisalong," or "barber shop."

In the past, language models were built from dictionaries by hand. But such systems are incomplete and don't reflect how people actually use language. Because our language models are based on users' interactions with Google, they are more precise and comprehensive -- for example, they incorporate names, idioms, colloquial usage, and newly coined words not often found in dictionaries.

When building our models, we use billions of web documents and as much historical search data as we can, in order to have the most comprehensive understanding of language possible. We analyze how our users searched and how they revised their searches. By looking across the aggregated searches of many users, we can infer the relationships of words to each other.

Queries are not made in isolation -- analyzing a single search in the context of the searches before and after it helps us understand a searcher's intent and make inferences. Also, by analyzing how users modify their searches, we've learned related words, variant grammatical forms, spelling corrections, and the concepts behind users' information needs. (We're able to make these connections between searches using cookie IDs -- small pieces of data stored in visitors' browsers that allow us to distinguish different users. To understand how cookies work, watch this video.)

To provide more relevant search results, Google is constantly developing new techniques for language modeling and building better models. One element in building better language models is using more data collected over longer periods of time. In languages with many documents and users, such as English, our language models allow us to improve results deep into the "long tail" of searches, learning about rare usages. However, for languages with fewer users and fewer documents on the web, building language models can be a challenge. For those languages we need to work with longer periods of data to build our models. For example, it takes more than a year of searches in Catalan to provide a comparable amount of data as a single day of searching in English; for Estonian, more than two and a half years worth of searching is needed to match a day of English. Having longer periods of data enables us to improve search for these less commonly used languages.

At Google, we want to ensure that we can help users everywhere find the things they're looking for; providing accurate, relevant results for searches in all languages worldwide is core to Google's mission. Building extensive models of historical usage in every language we can, especially when there are few users, is an essential piece of making search work for everyone, everywhere.