Personal discrimination on the Web

May 21 2009

How do you tell if a website you are browsing is a showing you a personal web page expressing the opinions of an individual or the marketing speak of a commercial site in disguise? Information engineers in India and Japan believe they have found an automatic way to discriminate between personal web pages and commercial pages designed to fool consumers.

Writing in a forthcoming issue of the *International Journal of Business Intelligence and Data Mining*, Takahiro Hayashi of Niigata University, and colleagues, explain that their approach extracts subjective expressions from web pages. The system then scores them by degree of subjectivity and provides the reader with an indication of whether the website content expresses personal opinions or marketing speak about a product or service.

The team has evaluated the performance of their system using 1200 web pages collected from four categories: product, tourist spot, restaurant, and movie. They found that their method is much more effective in finding personal opinion pages than a general search engine, in all categories. Part of the reason for this is that search engines, such as Google, tend not to rank personal pages highly.

Personal homepages, personal blogs, web forum sites and smaller customer opinion sites are regarded as personal pages and generally don't appear high in the search engine results pages (SERPs). Finding genuine personal opinions surveys is much harder than finding commercially biased sites, the researchers explain.

Their system relies on the fact that marketing copywriters and advertisers tend not to report negative comments about a product or service. In contrast, the personal opinions of users of the product or service will be littered with both positive and negative comments depending on their standpoint.

In Japanese, subjective expressions in written language might be described as: expressions with a negative meaning, sentence-final particles, interjections, and specific symbols such as face marks (Kanji), which are equivalent to smilies in the West. There are of course, equivalent expressions in other languages, say the researchers.

These various types of expressions can be extracted from a webpage and fed into the researchers' algorithm, which determines a weighted and categorized ratio of negative to positive expressions. This provides the basic indicator of whether or not a page is commercial or personal automatically.

The obvious extension of this approach is to apply such an algorithm to the results of a search for a product or service carried out by a general search engine and so filter out the commercial from the personal and allow consumers to assess the wider opinions of the web community on that product.

More information: Discrimination of personal web pages by extracting subjective expressions" in *Int. J. Business Intelligence and Data Mining*, 2009, 4, 62-77

Source: Inderscience Publishers (news : web)