Wiki ranking: Bayesian statistics can score Wikipedia entries

August 6 2014

Wikipedia the free, online collaborative encyclopedia is an important source of information. However, while the team of volunteer editors endeavors to maintain high standards, there are occasionally problems with the veracity of content, deliberate vandalism and incomplete entries. Writing in the *International Journal of Information Quality*, computer scientists in China have devised a software algorithm that can automatically check a particular entry and rank it according to quality.

Jingyu Han and Kejia Chen of Nanjing University of Posts and Telecommunications, explain that the quality of data on Wikipedia has for many years been the focus of user attention. Its detractors suggest that it can never be a valid information source in the way that a proprietary encyclopedia might be because the contributors and editors are not under the direct control of a single publisher with a vested interest in quality control. Its supporters suggest that the social nature of contributions and edits and the online tracking of changes is one of Wikipedia's greatest strengths rather than a weakness.

Nevertheless, it would quiet the detractors if there were a way to quantify the quality of Wikipedia entries in an objective and automated manner. Now, Han and Chen have turned to Bayesian statistics to help them create just such a system. The notion of finding evidence based on an analysis of probabilities was first described by 18th Century mathematician and theologian Thomas Bayes. Bayesian probabilities were then utilized by Pierre-Simon Laplace to pioneer a new statistical method. Today, Bayesian analysis is commonly used to assess the

content of emails and to determine the probability that the content is spam, junk mail, and so filter it from the user's inbox if the probability is high.

Han and Chen have now used dynamic Bayesian network (DBN) to analyze in a similar manner the content of Wikipedia entries. They apply multivariate Gaussian distribution modeling to the DBN analysis, which gives them a distribution of the quality of each article so that entries might be ranked. Very low-ranking entries might be flagged for editorial attention to raise the quality. By contrast, high-ranking entries could be marked in some way as the definitive entry so that such an entry is not subsequently overwritten with lower quality information.

The team has tested its algorithm on sets of several hundred articles comparing the automated quality assessment by the computer with assessment by a human user. Their algorithm out-performs a human user by up to 23 percent in correctly classifying the quality rank of a given article in the set, the team reports. The use of a computerized system to provide a quality standard for Wikipedia entries would avoid the subjective need to have people classify each entry. It could thus improve the standard as well as provide a basis for an improved reputation for the online encyclopedia.

More information: Han, J. and Chen, K. (2014) 'Ranking Wikipedia article's data quality by learning dimension distributions', *Int. J. Information Quality*, Vol. 3, No. 3, pp.207.

Provided by Inderscience Publishers

https://phys.org/news/2014-08-wiki-bayesian-statistics-score-wikipedia.html