

Weblog Data Mining for Business Intelligence

Asst Prof Sun Aixin

Division of Information Systems

School of Computer Engineering, NTU

Weblogs are online personal diaries managed by easy-to-use software packages that allow single-click publishing of daily entries [1]. The contents of Weblogs include news, observations, views and discussions on various topics. According to the surveys by Pew Internet & American Life Project [2], 27% of Internet users read Weblogs and 12% of Internet users have posted comments or other material on Weblogs.

With the increasing popularity, Weblogs have become important information sources. Different from other information sources, Weblogs are frequently updated by individuals to express their own feelings or opinions. In this project, we would like to investigate how Weblogs data can be mined for business and marketing intelligence. For example, after a new product is launched, many users of the product may post their comments about the product on their blogs. Those timely feedbacks from individual users could be very useful for the business operator to reconsider the business strategy or further improve of the product. However, to get meaningful feedbacks from Weblogs, one needs to (i) monitor millions of Weblogs to determine whether a post is relevant to the product, (ii) extract paragraphs/sentences related to various aspects of the product such as the design, the features and the usability, and (iii) organize the extracted feedbacks into topical categories for easy browsing and searching.

The expecting outcomes of the project include novel techniques for large-scale Weblogs mentoring, paragraph/sentence-level information retrieval/extraction and classification.

Reference:

- [1] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information Diffusion Through Blogspace. In Proceedings of WWW. Pages 491—501. New York, USA. 2004.
- [2] L. Rainie. The State of Blogging. Pew Internet & American Life Project. Available at: http://www.pewinternet.org/PPF/r/144/report_display.asp. 2005