Beyond Lift in Predictive Modeling: Duo-mining to Glean Strategic Insights

TREO Talk Paper

Q B. Chung
Villanova University
q.chung@villanova.edu

Abstract

Coined by Tan *et al.* (2006), "duo-mining" recognizes the synergy produced by combining the traditional *data mining* techniques with the relatively new analytical methods called *text mining*, the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources (Hearst, n.d.). While data mining proves to be effective when the data to be analyzed is structured, attention given to text mining lately is on the rise along with the staggering growth in volume of textual data found in the public domain such as tweets and various customer reviews.

Making a case for duo-mining has been done through measuring the lift in predictive modeling. Data mining is known to produce a significant lift in comparison to the baseline predictive methods such as random walk. Use of text mining also shows to create a significant lift over the efficacy of baseline prediction. When the two methods are combined, the lift is even higher than either text mining or text mining is employed alone.

Customer reviews have proven to be a fertile ground to authenticate the efficacy of duo-mining by combining numerical data (star ratings) and textual data (customer review narratives). Customer reviews in popular e-commerce sites are intended to provide new or potential customers with helpful information to assist in purchase decisions (Mudambi & Schuff, 2010). However, helpfulness of such reviews suffers from the disconnect between the quantitative measures and the qualitative assessment, and studies have pointed out the disconnect between the two (Mudambi, Schuff, & Zhang, 2014; Chung 2018). The potential business value of enhanced predictive power gained through duo-mining notwithstanding, such a gain tends to be ephemeral and confined to narrow domains due to the inherent nature of data collected from social network sites.

In this TREO talk, we present a novel approach of applying the duo-mining techniques to proprietary data sets of historical college admission applications, whereby not only the lift gain in the predictive power is confirmed but also strategic insights are gleaned by noting the similarities and differences in essays and recommendations, following the *cluster hypothesis* (van Rijsbergen, 1979).

References

Chung, Q. 2018. "Gaining Lift in Customer Review Helpfulness through Duo-mining," *Proceedings of the NEDSI 2018 Annual Conference.*

Fan, W., Wallace, L., Rich, S., and Zhang, Z. 2006. "Tapping the Power of Text Mining," *Communications of the ACM* (49:9), pp. 77-82.

Hearst, M. "What is Text Mining?" http://people.ischool.berkeley.edu/~hearst/text-mining.html.

Mudambi, S. M., Schuff, D. & Zhang, Z. 2014. "Why Aren't the Stars Aligned? An Analysis of Online Review Content and Star Ratings," *Proceedings of the 47th Hawaii International Conference on System Sciences*, pp. 3139-3147.

Mudambi, S. M., Schuff, D. 2010. "What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com." *MIS Quarterly* (34:1), pp. 185-200.

van Rijsbergen, C. J. 1979. "Information Retrieval", (2nd ed.) Butterworth-Heinemann.