Coloring Right

Chinmay Kulkarni

1 Abstract

Is it possible to automatically create a color palette relevant to a topic? Could such a palette be used to guide color choices while visualizing data? Prior work as shown that when data is presented in colors that are relevant to the data-topic, it leads to better and faster understanding.

This project explores the possibility of automatically creating topic-relevant palettes for a large class of topics. This automated system finds color palettes based on a corpus of images (from either ImageNet, Flickr or Google Images). The hypothesis is that generating palettes based on images in these corpora that are labeled with the topic will be topic-relevant. While we haven't narrowed down on the exact technique to extract palettes from color pixel values, clustering and topic models seem promising.

The system will be evaluated on three metrics: how well users like the generated palettes, how topic-relevant they are perceived to be, and if they affect data understanding. For all three metrics, the algorithmically generated palettes will be compared against a randomly generated palette, and one generated by experts. For the likability and understanding metrics, the random palette will be chosen from the set of palettes generated for other topics by our system (so that only the relevance, not the base quality of the colors is considered). Likability will be measured using a Likert scale. For relevance, an association task is used: given a topic (e.g "US Politics") and one of the topic terms (e.g. "Democrat"), the participant chooses which color, among a set of displayed swatches, is relevant to it. For understanding, users will be shown differently-colored infographics, and participants will be timed while they answer conceptual questions related to the infographic. Since the three metrics may interact strongly, they will be studied in a within-subjects design. Participants will be drawn from workers on Mechanical Turk, and non-crowd study participants will be recruited through the class and through dorm mailing lists.

For this evaluation, a number of topics will be used; since the performance of the system is likely dependent on the topics chosen, this may affect the results. The current plan is to use topics from visualizations in the New York Times in 2009-2010, or the Many Eyes website (categorical visualizations).

2 Stakeholders

This project aims to provide color recommendations to non-experts. For example, stakeholders might be citizens who access freely available government data, or amateur analysts. Expert graphic designers or visualization experts are not considered stakeholders. Stakeholders will use the recommender as part of a larger visualization toolkit, such as Data Wrangler. Since creating a user-interface best suited for this recommender is outside the scope of this project, I cannot describe how exactly each stakeholder would use this system.

3 Related work

Prior work exists on automatic creation of color palettes. This work falls broadly in two categories. The first focuses on finding representative colors from images, that can be used as color palettes. The most recent of these is [8]. This line of research has so far focused only on extracting colors from a single image. This project extends this work by extracting colors from multiple, related images. I believe that some of the techniques used by [8], such as a weighted histogram that uses color saturation and neighborhood color coherence, can be adapted for multiple images too. Depending on constraints of time, I plan to explore some of these techniques.

The second category of research on palette generation focuses on optimizing visual properties, such as color saliency and perceptive color distance, both manual or rule-based, as pioneered by Brewer [3]; and with varying degrees of automation [6, 9]. I believe most such optimization research is complementary to this project, and can be used as a post-extraction step to optimize the colors chosen. Statistical work on color saliency is valuable, even if it hasn't been directly applied as a optimization objective; color saliency in the context outside data-visualization in [4, 1].

Topic models have been shown to be effective in information retrieval. Latent semantic analysis (and later, LDA), for instance, has been used to find "latent" similarities between concepts [5, 2]. Similar similarity-measures

have been computed for nodes in a graph [7]. While these similarity measures may help to better cluster color-values, they don't target the domain of color recommendations directly.

References

[1] R. Benavente, F. Tous, R. Baldrich, and M. Vanrell. Statistical modelling of a colour naming space. In *Proceedings of the 1st European Conference on Colour in Graphics, Imaging, and Vision (CGIV2002)*, pages 406–411.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] C. Brewer. Color use guidelines for data representation. In *Proceedings of the Section on Statistical Graphics, American Statistical Association*, pages 55–60, 1999.

[4] J. Chuang, M. Stone, P. Hanrahan, and S. Consulting. A Probabilistic Model of the Categorical Association Between Colors. 2008.

[5] S. Dumais, G. Furnas, T. Landauer, S. Deerwester, and R. Harshman. Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 281–285. ACM, 1988.

[6] C. Healey. Choosing effective colours for data visualization. In *Visualization'96. Proceedings.*, pages 263–270. IEEE, 1996.

[7] G. Jeh and J. Widom. SimRank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543. ACM, 2002.

[8] B. Morse, D. Thornton, Q. Xia, and J. Uibel. Image-Based Color Schemes. In *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, volume 3, pages III–497. IEEE, 2007.

[9] A. Zeileis, K. Hornik, and P. Murrell. Escaping RGBland: Selecting colors for statistical graphics. *Computational Statistics & Data Analysis*, 53:3259–3270, 2009.