OHSU @ MediaEval 2015: Adapting Textual Techniques to Multimedia Search

Shiran Dudy
Center for Spoken Language Understanding
OHSU, Portland, Oregon
dudy@ohsu.edu

Steven Bedrick
Center for Spoken Language Understanding
OHSU, Portland, Oregon
bedricks@ohsu.edu

ABSTRACT

In this paper, we present the motivation, process, results and analysis of results that we have worked on as part of our participation in the 2015 MediaEval Retrieving Diverse Social Images Task. This year, we adapted a recently-published technique for result diversification ("Relational Learning-to-Rank" [13]), borrowed from the world of standard document retrieval. As compared to the original work, our version makes certain changes to the ranking and comparison algorithm, and explores a variety of feature combinations specific to an image retrieval context. The key idea behind our technique was a greedy iterative approach to ranking search results, which attempted to balance relevance with redundancy by comparing candidate results to those already selected by the algorithm. Our approach worked tolerably well on many queries, but there is clearly room for improvement.

1 Introduction

Imagine you are in Munich, and it's the time of year when everybody is talking about Oktoberfest. Being unfamiliar with this festival, you perform an image search, to try and find out whether you'd like the event, and to discover what to expect should you attend. Unfortunately, your results consist of two hundred very similar images, all of the inside of beer tents. While certainly *relevant*, these results only show a small slice of what Oktoberfest is about: where are the parades, the concerts, the fairgrounds? A more *diverse* set of search results would have been much more useful in this situation.

The Retrieving Diverse Social Images task at the 2015 MediaEval workshop required participants to provide the most diverse and relevant images given a search query like "Oktoberfest." The organizers provided a detailed task description along with data set for development and evaluation, described fully in [6]. Our team chose to adapt a recent technique for search result diversification [13] that adapts "traditional" learning-to-rank methods [8] to incorporate diversity into its loss function.

2 Related Work

Search result diversification is a very active area of research in information retrieval. In principle, the general problem of identifying an optimal ranking that balances both relevance and diversity has been shown to be NP-complete [1], which means that most techniques rely on approximations of one kind or another. One common family of approximations descend from the greedy and iterative Maximal and Marginal Relevance approach [3], in which each successive document is chosen based on its similarity to the user's query and its dissimilarity to the set of already-chosen documents.

Another family of approaches directly model attributes of the query and of the documents, and then identify sub-sets of results that are representative of different combinations of attributes. For example, Agrawal et al. [1] use a taxonomy to model documents and queries, and identify a set of results that thoroughly covers the entries represented by the retrieved documents.

The approach our group used in this year's MediaEval fuses elements of both families of techniques. It is based on a paper by Zhu et al. [13] that describes an extension of Learning-to-Rank (LtR). Traditional LtR consists of learning a ranking function that attempts to assign a rank to a particular document given a particular query. Zhu et al.'s extension, "Relational Learning-to-Rank" (R-LtR), models result ranking as a sequential selection process, and their formulation incorporates knowledge about not only the document in question and the query, but also the set of documents that have already been selected.

3 Methodology

For a complete description of R-LtR, we refer the reader to the original paper [13]. In brief, R-LtR is an iterative scoring method that takes into account both the relevance of a textual document along with information about how similar it is to documents that have already been chosen. The algorithm represents documents as arbitrary feature vectors. Each successive document is scored against the documents that have already been chosen according to the following scoring function (equation 2 in [13]):

$$f_s(x_i, R_i) = w_r^T \mathbf{x_i} + w_d^T h_s(R_i), \forall x_i \in X \backslash S \qquad (1)$$

This scoring function combines information on relevance and diversity given the candidate document $\mathbf{x_i}$ (represented as a k-dimensional feature vector) and its "diversity matrix" R_i. This matrix is actually a "slice" of a three-way tensor mapping documents to documents along features; each value R_{ijk} represents the relationship between documents i and j in terms of feature k. For example, if we were to use the Jaccard similarity metric as our first feature, $R_{i,j,1}$ would consist of the Jaccard similarity of documents $\mathbf{x_i}$ and $\mathbf{x_j}$. This formulation allows us to combine entirely arbitrary features and relational functions.

Note further that R_i in equation 1 is defined as including all documents $x_j \in S$, where S is the set of documents that have already been chosen out of the set of all possible docu-

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

ments, X. In other words, R_i contains information relating document $\mathbf{x_i}$ to the already-selected documents. $X \backslash S$ refers to the remaining set of not-yet-selected documents. $h_s(R_i)$ refers to a relational function comparing document $\mathbf{x_i}$ to the entire set of documents in S.[1]

Finally, w_r and w_d are weight vectors corresponding to the relative weights of relevance and diversity, respectively. Equation 1 is used for prediction (i.e., scoring); Zhu et al. outline a training process that uses stochastic gradient descent to learn learn values for w_r and w_d. For reasons of space, we will not discuss training in this paper, and refer the reader to the full description in [13]. Note that, unlike Zhu et al., in this year's task we were given results that are already sorted in terms of "relevance" (according to Flickr's search engine). As such, we were able to simplify the precise algorithm described by Zhu et al., as we were able to use this existing relevance information instead of computing our own from scratch.

In order to adapt this algorithm to the image search domain, we identified combinations of features and appropriate distance metrics based on the shared task data. We represented "textual" information by transforming each image's "tags" and "description" features into a tf-idf-weighted bag-of-words representation, which we then processed using Latent Semantic Analysis (LSA) [4] to reduce its dimensionality. We also performed Latent Dirichlet Allocation (LDA) [2] on the tag/description data, in order to attempt to represent topic groups within the results. We computed similarity for these features using L^2 (Euclidean) distance; both feature sets were computed using the Gensim package.[2]

In addition to the textual features, we utilized several of the visual features provided by the shared task. Along with their distance metrics, we used "csd" (L^2) [10], "hog" (Bhattacharyya distance) [11], "cn" (L^2) [7], "cm" (Canberra distance) [5], "lbp" (χ^2) [12], and "glr" (L^1 Manhattan) [9]. All features were normalized such that larger values for the distance functions represented either higher degrees of similarity or higher degrees of diversity (for the values in R). For our run including user credibility data, we included "visualScore", "faceproportion", "tagSpecificity", "uniqueTags", and "locationSimilarity".

4 Submitted Runs

We trained four different models. The first, `run 1`, used only image (visual) features. `run 2` used the textual features described above (LSA and LDA on descriptions and tags). `run 4` and `run 5` combined both image and textual features with user credibility information. The textual features remained the same across runs; runs 4 and 5 experimented with using global image features (i.e., calculated on the entire image) versus features computed locally on image quadrants.

5 Results & Discussion

	all			*multi*			*single*		
	F	CR	P	F	CR	P	F	CR	P
run 1	0.46	0.40	0.60	0.47	0.41	0.60	0.44	0.36	0.59
run 2	0.42	0.33	0.66	0.45	0.35	0.72	0.38	0.30	0.60
run 4	0.46	0.39	0.60	0.47	0.41	0.60	0.44	0.36	0.59
run 5	0.41	0.30	0.67	0.42	0.32	0.73	0.37	0.29	0.60

Table 1: Test set results for all runs. "multi" and "single" refer to multi- vs. single-topic queries. Metrics are reported at the official cutoff of @20.

Our results are summarized in Table 1. Our visual-feature-only run (run 1) outperformed our text-feature-only run (run 2) in terms of Cluster Recall @ 20, but interestingly, not in terms of Precision @ 20. Incorporating textual and user information (run 4) did not seem to substantially alter our system's results as compared with our visual-only run, while changing from global to local visual features *did* have a small effect— gaining precision at a cost of diversity.

While the raw numbers did not vary spectacularly between runs, the query-level performance of the different feature sets did vary a great deal. Consider the query "concerts in Bucharest." This was our text-only system's best performing query in terms of CR@20, with a score of 0.71. In contrast, our visual-only run scored considerably lower in CR@20, with a score of 0.57. However, if we examine the actual rankings produced by the system, we notice that the text-only system's images all look quite *visually* similar: they are variations of a musician on stage in front of a microphone. Importantly, they are all *different* musicians at *different* concerts. Our visual-only system's results, on the other hand, included several images of the concert hall itself, as well as images of several non-musical events taking place on stage. Many of these results were not considered relevant by the judges, which means that our image-only system's precision suffered on this query. However, this query illustrates the behavior of the R-LtR algorithm, as well as its sensitivity to feature selection.

Another instructive query was "Amsterdam gay parade." Our visual-only run substantially outperformed our text-only run in terms of CR@20 (0.67 vs 0.20), although both enjoyed very high precision. Inspection of the results reveals that, indeed, the visual-only run included a wide variety of scene and view types; the text-only run's results were largely of specific floats, taken by attendees. The descriptions themselves tended to name the particular group or float pictured, but the images themselves were visually homogeneous. These examples clearly illustrate that some queries' notions of "diversity" are better captured by textual features than by visual features, and vice versa.

We note that our system performed consistently better on the multi-concept queries (rows labeled "m" in Table 1) than on the single-concept queries (rows labeled "s"). The difference was not large, but its consistency is notable. One possible explanation for this finding is that the multi-concept queries' results may contain more diversity in terms of visual and textual features than did the single-concept queries, which in turn gave our our R-LtR implementation additional information to use in making its ranking decisions.

6 Conclusions

Our adaptation of R-LtR to an image retrieval task shows that this approach to result diversification can work with a wide variety of features and distance metrics. Our results are promising, though clearly much work remains to be done in terms of feature engineering and parameter tuning. We also hope to extend the algorithm to include more adaptable feature weight vectors, to enable the system to give different weight to certain features (e.g., textual or visual feature subsets) depending on query or image characteristics. R-LtR is a flexible and powerful platform from which to begin such experiments.

[1] Zhu et al. propose several different methods of combining the data stored in R_i: taking the minimal distance (i.e., for all features k, taking $\min_{x_j \in S} R_{ijk}$), averaging, or taking the maximum distance.

[2] `http://radimrehurek.com/gensim/`

7 References

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 5–14, New York, New York, USA, Feb. 2009. ACM.

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

[3] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, New York, New York, USA, Aug. 1998. ACM Request Permissions.

[4] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

[5] Z.-C. Huang, P. P. Chan, W. W. Ng, and D. S. Yeung. Content-based image retrieval using color moment and gabor texture feature. In *2010 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 2, pages 719–724. IEEE, 2010.

[6] B. Ionescu, A. L. Gınsca, B. Boteanu, A. Popescu, M. Lupu, and H. Müller. Retrieving diverse social images at mediaeval 2015: Challenge, dataset and evaluation. In *MediaEval 2015 Workshop, Wurzen, Germany*, 2015.

[7] H. Y. Lee, H. K. Lee, and Y. H. Ha. Spatial color descriptor for image retrieval and video segmentation. *IEEE Transactions on Multimedia*, 5(3):358–367, 2003.

[8] T.-Y. Liu. *Learning to rank for information retrieval.* Springer, New York, 1st. ed edition, 2011.

[9] S. Selvarajah and S. Kodituwakku. Analysis and comparison of texture features for content based image retrieval. *International Journal of Latest Trends in Computing*, 2(1), 2011.

[10] T. Sikora. The mpeg-7 visual standard for content description-an overview. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):696–702, 2001.

[11] Z. Yin and R. Collins. Object tracking and detection after occlusion via numerical hybrid local and global mode-seeking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[12] G. Zhang, X. Huang, S. Z. Li, Y. Wang, and X. Wu. Boosting local binary pattern (lbp)-based face recognition. In *Advances in biometric person authentication*, pages 179–186. Springer, 2005.

[13] Y. Zhu, Y. Lan, J. Guo, X. Cheng, and S. Niu. Learning for search result diversification. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 293–302. ACM, 2014.