Using Canonical Correlation Analysis for Generalized Sentiment Analysis, Product Recommendation and Search

Siamak Faridani
Automation Sciences Lab
Department of Industrial Engineering and Operations Research
University of California, Berkeley
faridani@berkeley.edu

ABSTRACT

Standard Sentiment Analysis applies Natural Language Processing methods to assess an "approval" value of a given text, categorizing it into "negative", "neutral", or "positive" or on a linear scale. Sentiment Analysis can be used to infer ratings values for users based on textual reviews of items such as books, films, or products. We propose an approach to generalizing the concept to multiple dimensions to estimate user ratings along multiple axes such as "service", "price" and "value". We use Canonical Correlation Analysis (CCA) and derive a mathematical model that can be used as a multivariate regression tool. This model has a number of valuable properties: it can be trained offline and used efficiently on live stream of texts like blogs and tweets, can be used for visualization and data clustering and labeling, and finally it can potentially be incorporated into natural language product search algorithms. At the end we propose an evaluation procedure that can be used on live data when a ground truth is not available. Based on this model we present our preliminary results from empirical data that we have collected from our system Opinion Space. We show that for this dataset the CCA model outperforms the PCA that was originally used in Opinion Space.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Theory

Keywords

Canonical Correlation Analysis, NLP, Information Clustering and Labeling, Generalized Sentiment Analysis

RecSys'11, October 23–27, 2011, Chicago, Illinois, USA.
Copyright 2011 ACM 978-1-4503-0683-6/11/10 ...$10.00.

1. INTRODUCTION

Product reviews on websites sometimes allow for ratings on a number of different dimensions. For example the online shoes and clothing store, Zappos[1], allows customers to review each pair of shoes on six numerical dimensions (comfort, style, size, width, arch support and overall). Similarly TripAvisor[2], a website for reviews and advice on hotels and flights, allows users to rate each hotel on six dimensions (value, rooms, location, cleanliness, service and sleep quality). In addition to these numerical values, each reviewer provides a textual review of the product or service. In traditional approaches to recommender systems these textual reviews are sometimes ignored because of the complexity that they introduce to the models. In this paper we utilize these numerical and textual feedbacks to train our model. A new user can then express the properties of her desired product either in textual form or on numerical scales. We show that our model is capable of using either of these sets of inputs to come up with a set of product recommendations for the user.

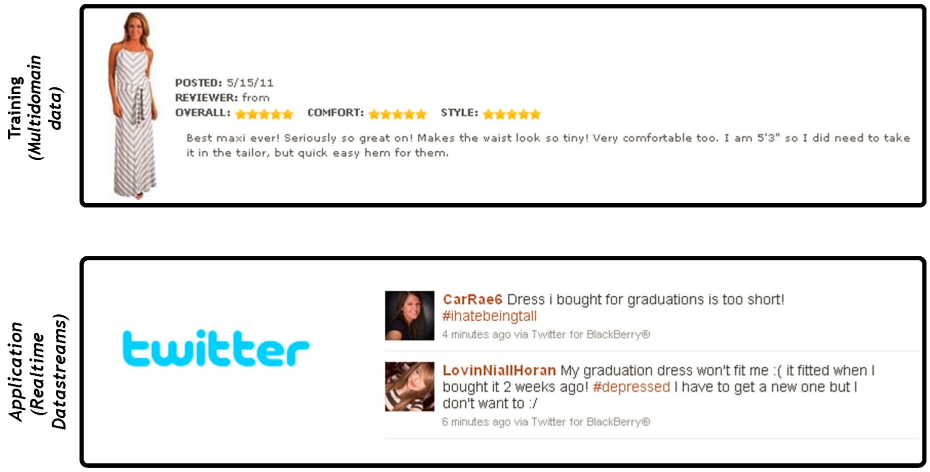
We use Canonical Correlation Analysis (CCA) to perform an offline learning on corpuses that have similar structures to Zappos and TripAdvisor in that they provide both textual reviews and numerical ratings. CCA is often used when two sets of data (x and y) are present and some underlying correlation is believed to exist between the two sets [8]. In our model x is the featurized representation of the textual review (i.e. an N-gram or a tf-idf representation of the text) and y is the vector of numerical ratings for each review. We hypothesize that combining both texts and numerical values enriches the recommendations and training. By using the data collected from our system, Opinion Space[3], we have validated this hypothesis. We have also developed an evaluation framework that enables us to test our models on live data when a ground truth is not present.

The mathematical structure of CCA allows for separation of offline learning and online use. Expensive learning process can be done offline on the dataset and learned mappings can be performed efficiently on live data streams coming from twitter and blogs. Additionally CCA considers the interdependence of response variables and we use this property to design a sentiment analysis model. This capability that we call "Generalized Sentiment Analysis" can be used for predicting the sentiment and its strengths on a number of different dimensions even in cases that these dimensions are not independent from each other. Additionally CCA can give the variance of the predicted values on different scales giving a confidence value for each predicted value.

[1] http://www.zappos.com

[2] http://www.tripadvisor.com/

[3] http://www.state.gov/opinionspace/


Training (Multidomain data)
Application (Realtime Datastreams)
twitter

POSTED: 5/15/11
REVIEWER: from
OVERALL: ★★★★★ COMFORT: ★★★★★ STYLE: ★★★★★

Best maxi ever! Seriously so great on! Makes the waist look so tiny! Very comfortable too. I am 5'3" so I did need to take it in the tailor, but quick easy hem for them.

CarRae6 Dress i bought for graduations is too short! #ihatebeingtall
4 minutes ago via Twitter for BlackBerry®

LovinNiallHoran My graduation dress won't fit me :(it fitted when I bought it 2 weeks ago! #depressed I have to get a new one but I don't want to :/
6 minutes ago via Twitter for BlackBerry®

Figure 1: In the example above we can train the model on dress reviews that are available on Zappos and then use the trained model as a regression model to find the expected numerical ratings for the live text streams from twitter. This model can be combined with product recommender systems for twitter users, or used as a sentiment analysis tool on live data.

2. CANONICAL CORRELATION ANALYSIS (CCA)

Originally proposed by Hotelling as a multivariate analysis method [7], Canonical Correlation Analysis has been used for text processing [3], multivariate prediction [1, 9], data clustering [4], data visualization [10, 8], image retrieval and search [6].

Our model is based on the probabilistic interpretation of Canonical Correlation Analysis presented by Bach and Jordan [2]. In this section we provide a brief overview of this method. For consistency, we follow the notation presented in Hardoon et al [6]. We also demonstrate this method for the classical presentation of CCA but the method is easily extendable to Kernel CCA as Kernel CCA is using CCA on the dataset after projecting onto a higher dimension feature space.

Let us consider two multivariate random variables x and y with zero mean and assume that they are correlated (zero mean assumption can be relaxed as any feature vector can be transformed by subtracting the mean). Denote each sample observation of x as x_i. For n samples we form two vectors $S_x = (x_1, ..., x_n)$ and $S_y = (y_1, ..., y_n)$. We now consider two linear directions w_x and w_y but we add this assumption that they have the same column rank such that $\langle w_x, x\rangle$ and $\langle w_y, y\rangle$ are projections in the same dimension vector space. This is an essential assumption in our model as it provides a number of interesting capabilities to the model. Here $\langle w_x, x\rangle$ is the inner product of vectors w_x and x and equals $w_x^T x$ (and w_x^T is the transpose of w_x). We can now look at elements of S in the new coordinate system thus we will have

$$\begin{aligned} S_{x,w_x} &= (\langle w_x, x_1\rangle, ..., \langle w_x, x_n\rangle) \\ S_{y,w_y} &= (\langle w_y, y_1\rangle, ..., \langle w_y, y_n\rangle) \qquad (1) \end{aligned}$$

CCA seeks to maximize the correlation between S_{x,w_x} and S_{y,w_y} thus the goal is to find w_x and w_y such that the following objective function (ρ) is maximized.

$$\begin{aligned} \rho &= \max_{w_x, w_y} corr(S_{x,w_x}, S_{y,w_y}) \\ &= \max_{w_x, w_y} \frac{\langle S_{x,w_x}, S_{y,w_y}\rangle}{||S_{x,w_x}|| ||S_{y,w_y}||} \end{aligned}$$

Hardoon et al show that the optimization problem of finding w_x and w_y linear transformations can be reduced to a symmetric eigenproblem and can be solved by a Cholesky decomposition [6]. S_y and S_x in this case are now the canonical representations of x and y. Meaning that S_x can be considered as the expected value of some latent variable z given x, $E(z|x)$. Similarly S_y is the expected value of the latent variable z given y, $E(z|y)$.

2.1 Using w_x and w_y for combining text and numerical ratings

Let us now consider the reviews on websites like TripAdvisor and Zappos where each review is also accompanied by one or many numerical ratings. We assume that the textual review is correlated with the numerical ratings. For example in the case of hotel ratings if a customer is not satisfied with the quality of the room and has given a low rating to the hotel on that scale we expect them to express their dissatisfactions in their comment text as well. These correlations will be captured by training the w_x and w_y matrices. Therefore we assume that

$$E(z|x) \approx E(z|y) \qquad (2)$$

This assumption allows us to map text and numerical spaces to the canonical space or alternatively map the canonical space back onto the text and ratings space. In this paper we utilize this to construct a model to predict the numerical ratings, y, from the text, x, meaning that if we observe the textual review, then by $\langle w_x, x\rangle$ we get $E(z|x)$ and since we assume Equation (2) holds we get $E(z|y) = \langle w_x, x\rangle$ also from Equation (1) we have $E(z|y) = \langle w_y, y\rangle$ therefore:

$$E(y) = w_y^{-1} w_x^T x \qquad (3)$$

After finding w_x and w_y from the training set of texts and numerical ratings, Equation (3) allows us to find the expected values of ratings by observing only the textual reviews. We will show that this equations has useful properties for multivariate regression, which can be applied to generalized sentiment analysis, search, visualization, data clustering and labeling.

3. GENERALIZED SENTIMENT ANALYSIS

Sentiment Analysis is traditionally performed on one attribute of the target products. We extend the model by looking at many different dimensions of the product together. CCA provides a supervised learning model for extracting the attributes of products and services from textual reviews. Unlike univariate Support Vector Machine models (SVM), CCA allows us to consider the interdependence among response variables.

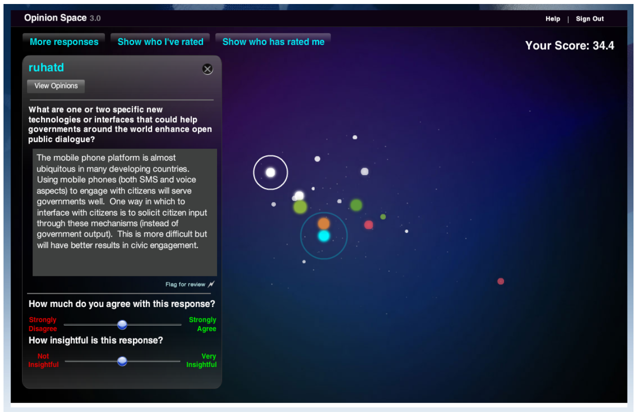
One aspect of the CCA model is that it can be trained on datasets like Zappos or TripAdvisor and then used online to extract the sentiment of the market from sources like blogs and tweeter feeds that lack the numerical value for reviews. It can also be used to highlight the key words that contribute to major changes in the numerical scales. We can calculate the effect of increasing the frequency of each word to the changes in each numerical scale. For example we hypothesize that words such as “comfortable” in “this shoe is comfortable” will cause major changes in the numerical value of the “comfort” scale while in the sentence “my uncle wears this shoe” the term “my uncle” will cause no change in the numerical value of comfortability. Also fitting parameters to the CCA model that is the most expensive part can be done offline. The online procedure which is multiplying the learned transformation matrices (w_x and w_y) with the featurized text can be done cheaply in real-time. Following applications are proposed for this model: filling the missing values for reviews (for example if the design of the website is changed and there is no numerical values for reviews before some certain time), inferring the expected ratings for unstructured reviews that are expressed outside the company’s website (for example if we observe a blog post that reviews a hotel we can use our CCA model to find the expected ratings associated with that post on different dimensions). Another application would be to have a textbox for users to enter their desired properties for their trip. For example something similar to the following: “I am looking for a hotel that is pet friendly, in a good neighborhood of the city and I don’t care about hotel amenities I just want it to be affordable”. By using Equation (3) on the featurized text. Our CCA model can then infer and extract numerical values for each dimension of the numerical scale and then by running a K-Nearest Neighbors algorithm, search on hotels that have the closest properties to the provided query. This can serve as part of a natural language search engine for products or a natural language product recommender engine.

4. TEST CASE

Neither Zappos nor TripAdvisor provide API access to their datasets. In this paper we use the data that is collected from our own system, Opinion Space, that is deployed at the US Department of State website and provides the same structure in the data.

4.1 Opinion Space as an evaluation platform

Opinion Space [5] is a collaborative tool for collecting insights and ideas on different challenges. It is being used by the US Department of State to collect ideas on US foreign policy and to promote discussion among individuals around the world[4]. The underlying dataset of Opinion Space is very similar to that of websites like Zappos and TripAdvisor. Each user provides one textual response to a question and they also provide numerical ratings on how much they agree or disagree with a number of statements. Additionally, participants rate each other’s responses based on how much they agree with them and how insightful they find those responses. We encourage interested readers to refer to the website and our past publications on Opinion Space for further details [5].



What are one or two specific new technologies or interfaces that could help governments around the world enhance open public dialogue?

The mobile phone platform is almost ubiquitous in many developing countries. Using mobile phones (both SMS and voice aspects) to engage with citizens will serve governments well. One way in which to interface with citizens is to solicit citizen input through these mechanisms (instead of government output). This is more difficult but will have better results in civic engagement.

How much do you agree with this response?

How insightful is this response?

Figure 2: *Opinion Space interface*. Each point in the space represents one participant. In the current version Principal Component Analysis (PCA) has been used to project user responses onto a 2D space. A user can click on different points and see a participant’s textual response to a discussion question. One can rate comment responses for participants on two scales, agreement and insightfulness. Comment ratings provided by participants help us develop recommendation systems to determine the most insightful ideas in the space.

Opinion Space collects opinions on baseline statements as scalar values on a continuous scale and applies dimensionality reduction to project the data onto a two dimensional plane for visualization and navigation (Figure 2). This technique effectively places all participants onto a level playing field. Points far apart correspond to participants with very different opinions while participants with similar opinions are proximal (the converses do not necessarily hold). Participants in Opinion Space contribute textual responses to discussion questions and are encouraged to earn points through reading and rating the responses of others.

With over 4,700 proposition ratings, more than 2,400 textual comments and over 17,400 numerical ratings for the comments this collection can serve as a data set for different natural language processing algorithms.

5. CLUSTER ANALYSIS AND REGION LABELS

One interesting aspect of CCA is that it provides a topic model. Recall that CCA gives linear transformations w_x and w_y that can embed two vectors (featurized comments and numerical values) to a lower dimensional canonical space. We can alternatively use w_x^{-1} to go from the canonical space to the text space. So each point in the canonical space will have a likelihood of each topic associated to it. By numerically integrating over a region we can find the main topic in that specific region. The following simple algorithm shows how we can use CCA to extract topics from the corpus.

6. EVALUATION AND PRELIMINARY RESULTS

For evaluating the effectiveness of our CCA method we developed the following evaluation algorithm. This algorithm

[4] www.state.gov/opinionspace/

```
Input: 2D projections in the canonical space $Z$, $w_x^{-1}$
Output: Topic labels for each region
Run a K-means on the list of comments and cluster them into $k$ clusters;
foreach cluster $c$ do
    initialize vector $r_c = \mathbf{0}$;
    foreach point in $Z$ do
        Calculate $r_c = r_c + w_x^{-1} Z$
    end
    Tag the cluster with the topic that has the maximum value in $r_c$
end
```

Algorithm 1: Labeling the regions in the canonical space by finding the topic with the maximum expectation in each cluster

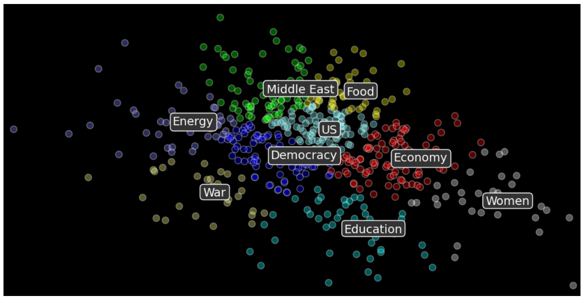


Figure 3: Cluster Analysis: CCA enables us to look at how participants' responses form clusters in the 2D space. Opinion Space 2.0 dataset is used for this analysis and participants provided responses on what would they tell secretary Clinton if they see her. Responses varied from issues about women rights to energy policies. We used the cluster analysis and the region labeling detailed in section . As we see CCA has placed topics like *Women* and *Education* near one another. Also *Middle-East* and *Energy* are also placed close to each other.

allows us to evaluate our models on live data where a ground truth does not exist. We utilize the ratings that individual users provide on each comment. We assume that the best dimensionality reduction should place reviews on an Euclidian space such that the ones that rated each other positively are placed closer to each other and the ones that have rated each other negatively are far from one another.

We have taken users of Opinion Space and their ratings and projected each user by the new method (CCA) and by PCA. We then looked at the Euclidean distance between each two users and their agreement rating. The correlation between these two values are shown in the following table. One of the fundamental assumptions in Opinion Space is that similar opinions are placed closer to each other. As we see, CCA provides the highest correlation and we conclude that by combining both the textual comments and numerical ratings CCA is a better dimensionality reduction method when compared to PCA (in this case PCA only considers numerical ratings).

DM Method	Pearson's Correlation
CCA	0.352
PCA	0.134
Random	0.002

Table 1: CCA provides a higher correlation between the agreement values between participants and the Euclidean distance between them. CCA was performed to combine the text and numerical inputs while PCA only considers the numerical ratings for each participant and ignores their textual inputs

7. NEXT STEPS

We are working on extending the CCA model to develop a robust CCA method in which outliers do not change the w_x and w_y mappings. Better featurization of the text also greatly influences the final performance of the model. So far, we have only focused on a bag-of-word representation of each document but a proper POS model or an N-gram model seems to be the natural extension for the featurization. Additionally, we are interested in techniques for reducing the amount of information stored in memory. We are planning to use point-wise mutual information (PMI-IR) to cluster similar words together and reduce the size of the large feature vectors for the text. More experimentations with other datasets are also included in our future directions for this research.

8. ACKNOWLEDGMENTS

The author thanks Ken Goldberg, Ephrat Bitton, David Wong, Taylor Berg-Kirkpatric, Timmy Siauw, Dmitry Berenson and Jun Wang for their valuable feedbacks on this work. This project is supported in part by the Berkeley Center for New Media, Fujitsu Labs of America and Berkeley AMPlab.

9. REFERENCES

[1] B. Abraham and G. Merola. Dimensionality reduction approach to multivariate prediction. *Computational statistics & data analysis*, 48(1):5–16, 2005.

[2] F. Bach and M. Jordan. A probabilistic interpretation of canonical correlation analysis. *University of California, Berkeley, Tech. Rep*, 2005.

[3] J. Blitzer, D. Foster, and S. Kakade. Domain adaptation with coupled subspaces.

[4] K. Chaudhuri, S. Kakade, K. Livescu, and K. Sridharan. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th annual international conference on machine learning*, pages 129–136. ACM, 2009.

[5] S. Faridani, E. Bitton, K. Ryokai, and K. Goldberg. Opinion space: a scalable tool for browsing online comments. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 1175–1184. ACM, 2010.

[6] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[7] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[8] P. Lai and C. Fyfe. A latent variable implementation of canonical correlation analysis for data visualisation. In *Neural Networks, 2006. IJCNN'06. International Joint Conference on*, pages 1143–1149. IEEE, 2006.

[9] P. Rai and H. Daumé III. Multi-label prediction via sparse infinite cca. *Advances in Neural Information Processing Systems*, 22:1518–1526, 2009.

[10] T. Sun and S. Chen. Locality preserving cca with applications to data visualization and pose estimation. *Image and Vision Computing*, 25(5):531–543, 2007.