Predicting Interest: Another Use for Latent Semantic Analysis

Thomas J. Connolly (connot3@rpi.edu)
Vladislav D. Veksler (vekslv@rpi.edu)
Wayne D. Gray (grayw@rpi.edu)
Department of Cognitive Science, Rensselaer Polytechnic Institute
110 8th Street
Troy, NY 12180 USA

Abstract

Latent Semantic Analysis (LSA) is a statistical technique for extracting semantic information from text corpora. LSA has been used with success to automatically grade student essays (Intelligent Essay Scoring), model human language learning, and model language comprehension. We examine how LSA may help to predict a reader's interest in a selection of news articles, based on their reported interest for other articles. The initial results are encouraging. LSA (using default corpus and setup) can closely match human preferences, with RMSE values as low as 2.09 (human ratings being on a scale of 1-10). Additionally, an Adapting Measure (best parameters for each individual) produced significantly better results, RMSE = 1.79.

Keywords: Adapting Measure; Latent Semantic Analysis; LSA; human interest prediction; predicting ratings; news articles

Introduction

The ability to accurately predict user preferences is valuable to any system that strives to deliver meaningful data based on a single query. It allows improved accuracy in the results delivered and by extension, superior service to a system's users. One example of this is the Netflix's Cinematch algorithm, which uses linear statistical models to provide their users with estimates of how much they will enjoy a certain movie given their prior movie preferences (Netflix, 1997). Latent Semantic Analysis (LSA; Landauer & Dumais, 1997), a technique designed to find relations between bodies of text, may offer a suitable alternative for rating text documents. LSA has been employed in a similar manner to predict grades for student essays (Foltz, Laham, & Landauer, 1999), making it worth exploring its worth as a model for user preferences. This paper examines LSA's capability to predict user preferences for news articles and outlines an experiment designed and used to this end. The problem space was defined by 3 factors. First, we examine the effect of using different amounts of the articles' content (title only versus title + content) on prediction accuracy. Second, we examine how different methods for predicting interest compare (e.g. average rating of 3 closest articles, weighted average of 9 closest ratings, etc.). Lastly, we evaluated nomothetic (one size fits all) and idiographic (tailored to individuals) approaches to predict user preferences.

Background

Latent Semantic Analysis Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) is a statistical technique for extracting semantic information from text corpora. It is a powerful technique that has been used with success for automatically grading student essays (Landauer & Dumais, 1997), to model human language learning (Landauer & Dumais, 1997), to model language comprehension (Lemaire, Denhiere, Bellissens, & Jhean-Iarose, 2006), and more.

Although this paper focuses on LSA, other techniques for modeling the human semantic space may be appropriate (Blei & Lafferty, 2006; Blei, Ng, & Jordan, 2003; Griffiths & Steyvers, 2004; Lindsey, Stipicevic, Veksler, & Gray, 2008; Matveeva, Levow, Garahat, & Royer, 2005; Veksler, Govostes, & Gray, 2008), and will be assessed in future work. Moreover, we examine only one of many possible LSA spaces, based on the work of Landauer & Dumais (1997), constructed based on the TASA corpus (Zeno, Ivens, Millard, & Duvvuri, 1995).

The TASA corpus contains a body of text which represents a collection or reading material that a college freshman should be familiar with (Zeno, Ivens, Millard, & Duvvuri, 1995). The LSA-TASA space has been used frequently as a model of human semantics (e.g. Griffiths, Steyvers, & Tenenbaum, 2007; Landauer & Dumais, 1997; Veksler, Govostes, and Gray, 2008;), and is an appropriate first model to use for current research. However, corpus selection makes a great difference for any semantic modeling (Lindsey, Veksler, Grintsvayg, and Gray, 2007), and other corpora will be employed in future research to further improve predictions of human ratings.

Intelligent Essay Scoring Intelligent Essay Scoring, in particular the Intelligent Essay Assessor (IEA) is relevant to current work. Intelligent Essay Assessment, in short, utilizes LSA to grade student essays by comparing them with essays of known quality based on the degree of conceptual relevance and the amount of relevant content. When put to the test, Results indicated that this technique varied from a human grader as much as two human graders varied from each other. This shows that the Intelligent Essay Assessor performed almost identically to human graders, showing a great deal of support for LSA as a measure of semantic similarity (Foltz, Laham, & Landauer, 1999). The proposed model differs from IEA in that it does not work with a predetermined set of ranked bodies of text, opting instead to learn the user's ranking system and attempting to emulate it.

Theory

LSA represents semantics as a multidimensional vector-space. A given body of text can be represented in this semantic space by a vector. The relatedness between any two

such vectors can then be measured based on the angle between them. The greater the angle between two vectors, the greater the difference semantically between the two concepts represented by said vectors.

We believe that humans assign utility values to semantic concepts, and that these values can be measured and utilized to model human interest. The assumption is that the closer two articles are in semantic space, the closer their interest values should be for any given participant. Figure 1 helps to illustrate this idea further, where similar semantic topics have similar interest values for a sample human participant. Thus, the interest value for any new vector drawn in this semantic space may be predicted based on which existing vectors it is closest to.

The experiment described below tests this prediction. The idea is that by comparing the articles having known human interest values with a new article having none yet assigned, we can predict the utility value of the new article. By taking the average interest value of the n most closely related LSA vectors, we can infer a value for the unknown article. Giving more weight to vectors that are more closely related to the unvalued article's vector may increase the prediction's accuracy (this may compensate for cases where the known semantic interest space is sparse, and only a small number of articles have a high relatedness to the unrated article).

Modeling

In the experiment we explore the problem space mentioned in the introduction based on its three defining factors.

Content Size

The experiment inspected the effect content size held over the accuracy of the predictions. When considering news articles, we used the title text versus the full article text to examine this. We wanted to know if the titles' of the articles alone would give enough information for rating predictions. The assumption was that the article titles would give a fair indication of the article's representation in a semantic space (as is the case much of the time for human readers). On the other hand it may be better to base predictions on the full set of data, in this case article content.

Measures

Several measures were used to predict utility values to each user's articles. Averages of the n closest related articles to the one being rated were used, with $n = 3,5,7,9,11,15,25,33,100,299$. Weighted averages and double weighted averages of the same amounts were also calculated in averages of the n closest vectors. The weighted average was used to determine a predicted interest for article a as follows:

$$WA(a) = \frac{\sum_{k=1}^{n} LSA(a,x_n) \times Rating(x_n)}{\sum_{k=1}^{n} LSA(a,x_n)} \quad (1)$$

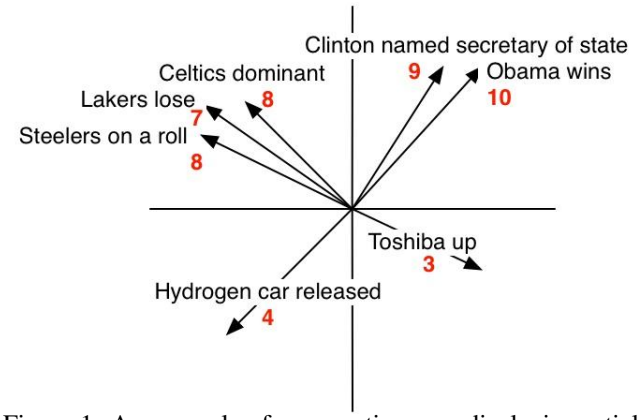


Figure 1: An example of a semantic space displaying article vectors, their titles, and the utility value assigned to them by a fictitious human participant.

where x_k is the nth most closely related article and $LSA(a,x_k)$ is the measure of the relatedness between a and x_k . The Double weighted average is calculated almost the same except that the measure of the relatedness between a and x_k is squared like so:

$$WA(a) = \frac{\sum_{k=1}^{n} LSA(a,x_n)^2 \times Rating(x_n)}{\sum_{k=1}^{n} LSA(a,x_n)^2} \quad (2)$$

Each measure was tested using the LSA vectors for only the article's title as well as the article's title and content. There is no comparison between articles rated by different users. Also, the averages were rounded, so that an integer value was assigned as the predicted rating. For each user, the root mean squared error (RMSE) of the predicted utility values from the user defined utility values was calculated as an indication of overall performance.

Nomothetic Versus Idiographic

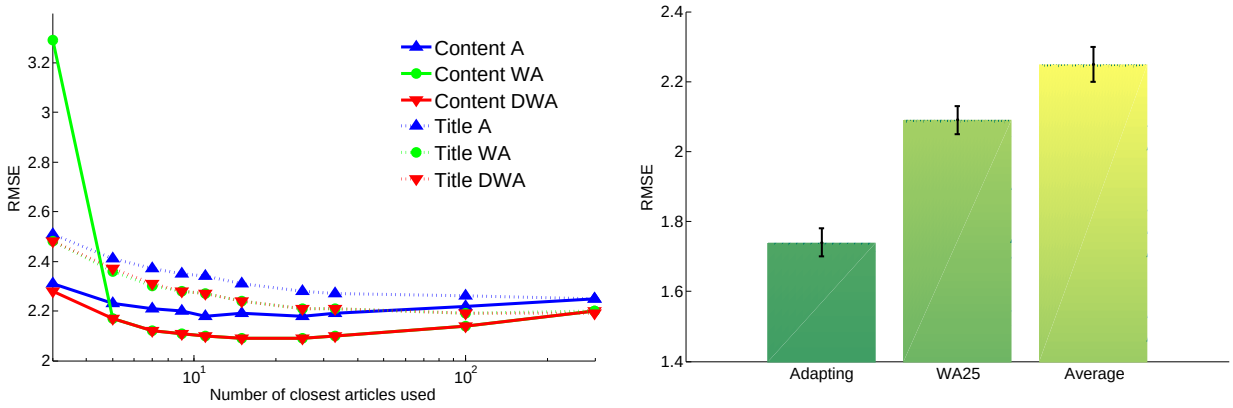
Finally the prediction accuracy between two distinct approaches was measured and compared. The Nomothetic approach simply used one content size and one measure with every user. This static approach was applied to every combination of content size and measure. The idiographic approach involved an Adapting Measure, which tried every combination of content size and measure to predict a given user's utility values, and chose the most accurate on a user by user basis.

Experiment

The primary purpose of this experiment was to measure the accuracy of LSA in predicting a user's interest in regards to news articles based off that user's previous ratings.

Procedure and Design

Participants 200 undergraduate students of RPI participated for course-credit. Twelve of the participants failed to finish the experiment, and their data was subsequently removed from any further analysis.



(a) (b)

Figure 2: **(a)** Average RMSE values for Average (A), Weighted Average (WA), and Double-Weighted Average (DWA) measures for predicting interest values, using either the article Title, or its Content to create the LSA vector. **(b)** A comparison of the average RMSE values of an Adapting Measure, the nomothetic measure with the lowest RMSE (weighted average of 25 articles), and the Average Interest Heuristic (average of 299 articles).

Design The experiment employed a single-group design with no between-subject variables. 400 news articles were taken from Reuters.com and classified into 20 categories based on their content. The categories used were: Sports, Entertainment, US News, Environment, Health News, Lifestyle-Health, Politics, International, Business News, Deals, Private Equity, Mergers & Acquisitions, Science, Internet, Lifestyle-Technology, Technology, Lifestyle-Travel, Oddly Enough, Lifestyle-Living, and Lifestyle-Autos. Each category contained 20 articles. These articles were pulled from the Reuters RSS feed the week before the experiment was run so as to offer the most recent articles possible to the Participants for grading. The experiment itself was designed as a web application and the Participants were instructed to complete it at home, allowing a greater number of participants to contribute.

Procedure Before the online experiment began, the participants were instructed to provide their name, gender, age, and major. Participants were then instructed to rate 300 articles, chosen randomly from the aforementioned set of 400 articles. Each article was to be rated on a scale of 1 to 10 with 1 signifying indifference and 10 signifying that an article perfectly matches the participants interest. The participants were also told that they did not have to read the articles if they could gauge their interest by the title alone. The experiment appeared in the browser as a list of 10 article titles hyperlinked to their source with 10 radio buttons underneath each title labeled 1-10 to allow the participants to submit their ratings with ease. Once the participants rated 10 articles, they would be able to click a button at the bottom of the screen which would reload the page with 10 new articles. The experiment would not allow the participants to move on to the next page of 10 articles without first rating the 10 that were currently displayed After the 300 articles were rated, the participants were then asked to complete a questionnaire that gave us valuable feedback in regards to the experiment's procedure. Pilot participants were able to finish the experiment in less than one hour.

Results and Analysis

Each measure's performance (i.e. how accurately they predicted the participants' ratings) is displayed as RMSE values in Figure 2 and Table 1.

	Content			Title		
n	**A**	**WA**	**DWA**	**A**	**WA**	**DWA**
3	2.31	3.29	2.28	2.51	2.48	2.48
5	2.23	2.17	2.17	2.41	2.36	2.37
7	2.21	2.12	2.12	2.37	2.30	2.31
9	2.20	2.11	2.11	2.35	2.28	2.28
11	2.18	2.10	2.10	2.34	2.27	2.27
15	2.19	2.09	2.09	2.31	2.24	2.24
25	2.18	**2.09**	2.09	2.28	2.21	2.21
33	2.19	2.10	2.10	2.27	2.21	2.21
100	2.22	2.14	2.14	2.26	2.19	2.19
299	**2.25**	2.20	2.20	2.25	2.20	2.19

Table 1: RMSE values from the graph in Figure 2a.

It appears that using the actual content of the article to fill the semantic space is superior to using just the title text. This is most likely due to the greatly increased amount of text used to create the article vectors. Larger bodies of text allow for stronger similarity between the articles' content, and therefore better results. Focusing purely on the content based information, it is evident that there is no benefit to using double weighted averaging, as it offers almost identical results to just using the weighted averages. The best overall measures seen here are the weighted averages of 15 and 25 articles (WA15 and WA25). Given this information we can say that the best nomothetic measures are weighted averages of somewhere between 15 and 25 of the most closely related articles.

Also worth noting is the performance increase from using the average rating of all 299 articles. Using the participant's average interest value as a heuristic for approximating their interest in any one article, results in an average RMSE of 2.25. This Average Interest Heuristic may be used as a performance baseline. The average difference in RMSE values between WA25 and the Average Interest Heuristic is .16. This is a dramatic difference, considering that real-world rating prediction algorithms are competitive to the RMSE values of .001. Consider, for example, the Netflix Prize contest where RMSE improvements in the thousandths place are the difference between being in the top 5 and being in the top 26. (Netflix, 1997).

Although WA25 produces the lowest average RMSE, greater accuracy can be achieved by using an Adapting Measure. By choosing the best measure for each participant, performance is increased. In other words, whereas WA25 may be the best rating predictor for one participant, a simple WA15 may be more appropriate for another. The average RMSE value for using the Adapting Measure is 1.74. A repeated measures ANOVA revealed significant differences between the Adapting Measure (M=1.74, SE=.04), WA25 (M=2.09, SE=.04), and the Average Interest Heuristic (M=2.25, SE=.05), $F(2, 557) = 38.272$, $p < .001$.

Conclusions

The experiment determined that LSA warrants further study as a model of predicting human interest. Initial results for predicting participants' interests in news articles (using the default LSA corpus and setup) were very positive, resulting in RMSE values as low as 2.09 using a nomothetic method. The idiographic method resulted in significantly better performance still, RMSE = 1.74. A greater degree of article content seems to lead to more informative LSA vectors, and better rating predictions. Lastly, we have narrowed down the list of measures for further examination to Weighted Averaging of 15 to 25 closest articles, disregarding Averaging and Double-Weighted Averaging methods of rating estimation. With further study and experimentation we believe that this impressive level of accuracy can be improved to an even greater precision.

Future experiments will involve rating predictions for more diverse text (e.g. comics, books, scientific papers). Modifications to the current model will be explored, using alternative (more modern) training corpora for LSA, and different modeling techniques.

Acknowledgements

The work was supported, in part, by grant N000140710033 to Wayne Gray from the Office of Naval Research, Dr. Ray Perez, Project Officer. We thank Robert Lindsey for his assistance.

References

Balabanovic, M. (1998). Exploring versus exploiting when learning user models for text representation. User Modeling and User-Adapted Interaction, 8, 71-102.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. The Journal of Machine Learning Research, 3, 993-1022.

Blei, D., & Lafferty, J. (2006). Correlated topic models. In Advances in neural information processing systems 18.

Foltz, P., Laham, D. & Landauer, T. (1999). Automated Essay Scoring: Applications to Educational Technology. In B. Collis & R. Oliver (Eds.), Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 1999 (pp. 939-944). Chesapeake, VA: AACE.

Golder, S. and Huberman, B. A. 2005 The structure of collaborative tagging systems. Technical report, In-formation Dynamics Lab, HP Labs.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences, 101, 52285235.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. Manuscript submitted for publication.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.

Lemaire, B., Denhiere, G., Bellissens, C., & Jhean-Iarose, S.(2006). A computational model for simulating text comprehension. Behavior Research Methods, 38(4), 628-637.

Lindsey, R., Veksler, V. D., Grintsvayg, A., & Gray, W. D. (2007). Effects of Corpus Selection on Semantic Relatedness. 8th International Conference of Cognitive Modeling, ICCM2007, Ann Arbor, MI.

Lindsey, R., Stipicevic, M., Veksler, V.D., & Gray, W. D. (2008). BLOSSOM: Best path Length On a Semantic Self-Organizing Map. 30th Annual Meeting of the Cognitive Science Society.

Matveeva, I., Levow, G., Farahat, A., & Royer, C. (2005).Term representation with generalized latent semantic analysis. Presented at the 2005 Conference on Recent Advances in Natural Language Processing.

Netflix. (1997-2009). Netflix Prize: View Leaderboard. Retrieved April 8, 2009, from http://www.netflixprize.com/leaderboard.

Netflix. (1997-2009). Netflix Prize: View FAQ. Retrieved April 12, 2009, from http://www.netflixprize.com/faq.

Paul Heymann, Georgia Koutrika, and Hector Garcia Molina, Can Social Bookmarking Improve Web Search?. Resource Shelf, (2008).

Veksler, V. D., Govostes, R. Z., & Gray, W. D. (2008). Defining the Dimensions of the Human Semantic Space. Proceedings of the 30th Annual Meeting of the Cognitive Science Society, CogSci08.