Predicting Clicks in a Vocabulary Learning System

Aaron Michelony
Baskin School of Engineering
University of California, Santa Cruz
1156 High Street
Santa Cruz, CA 95060
amichelo@soe.ucsc.edu

Abstract

We consider the problem of predicting which words a student will click in a vocabulary learning system. Often a language learner will find value in the ability to look up the meaning of an unknown word while reading an electronic document by clicking the word. Highlighting words likely to be unknown to a reader is attractive due to drawing his or her attention to it and indicating that information is available. However, this option is usually done manually in vocabulary systems and online encyclopedias such as Wikipedia. Furthurmore, it is never on a per-user basis. This paper presents an automated way of highlighting words likely to be unknown to the specific user. We present related work in search engine ranking, a description of the study used to collect click data, the experiment we performed using the random forest machine learning algorithm and finish with a discussion of future work.

1 Introduction

When reading an article one occasionally encounters an unknown word for which one would like the definition. For students learning or mastering a language, this can occur frequently. Using a computerized learning system, it is possible to highlight words with which one would expect students to struggle. The highlighting both draws attention to the word and indicates that information about it is available.

There are many applications of automatically highlighting unknown words. The first is, obviously, educational applications. Another application is foreign language acquisition. Traditionally learners of foreign languages have had to look up unknown words in a dictionary. For reading on the computer, unknown words are generally entered into an online dictionary, which can be time-consuming. The automated highlighting of words could also be applied in an online encyclopedia, such as Wikipedia. The proliferation of handheld computer devices for reading is another potential application, as some of these user interfaces may cause difficulty in the copying and pasting of a word into a dictionary. Given a finite amount of resources available to improve definitions for certain words, knowing which words are likely to be clicked will help. This can be used for caching.

In this paper, we explore applying machine learning algorithms to classifying clicks in a vocabulary learning system. The primary contribution of this work is to provide a list of features for machine learning algorithms and their correlation with clicks. We analyze how the different features correlate with different aspects of the vocabulary learning process.

2 Related Work

The previous work done in this area has mainly been in the area of predicting clicks for web search ranking. For search engine results, there have been several factors identified for why people click on certain results over others. One of the most important is position bias, which says that the presentation order affects the probability of a user clicking on a result. This is considered a “fundamental problem in click data” (Craswell et al., 2008), and eye-

tracking experiments (Joachims et al., 2005) have shown that click probability decays faster than examination probability.

There have been four hypotheses for how to model position bias:

- Baseline Hypothesis: There is no position bias. This may be useful for some applications but it does not fit with the data for how users click the top results.
- Mixture Hypothesis: Users click based on relevance or at random.
- Examination Hypothesis: Each result has a probability of being examined based on its position and will be clicked if it is both examined and relevant.
- Cascade Model: Users view search results from top to bottom and click on a result with a certain probability.

The cascade model has been shown to closely model the top-ranked results and the baseline model closely matches how users click at lower-ranked results (Craswell et al., 2008).

There has also been work done in predicting document keywords (Doğan and Lu, 2010). Their approach is similar in that they use machine learning to recognize words that are important to a document. Our goals are complimentary, in that they are trying to predict words that a user would use to search for a document and we are trying to predict words in a document that a user would want more information about. We revisit the comparison later in our discussion.

3 Data Description

To obtain click data, a study was conducted involving middle-school students, of which 157 were in the 7th grade and 17 were in the 8th grade. 90 students spoke Spanish as their primary language, 75 spoke English as their primary language, 8 spoke other languages and 1 was unknown. There were six documents for which we obtained click data. Each document was either about science or was a fable. The science documents contained more advanced vocabulary whereas the fables were primarily written for English language learners. In the study, the students took a vocabulary test, used the vocabulary system and then took another vocabulary test with the same words. The highlighted words were chosen by a computer program using latent semantic analysis (Deerwester et al., 1990) and those results were then manually edited by educators. The words were highlighted identically for each student. Importantly, only nouns were highlighted and only nouns were in the vocabulary test. When the student clicked on a highlighted word, they were shown definitions for the word along with four images showing the word in context. For example, if a student clicked on the word “crane” which had the word “flying” next to it, one of the images the student would see would be of a flying crane. From Figure 1 we see that there is a relation between the total number of words in a document and the number of clicks students made.

Number	Genre	Words	Students
1	Science	2935	60
2	Science	2084	138
3	Fable	667	23
4	Fable	513	22
5	Fable	397	16
6	Fable	105	5

Table 1. Document Information

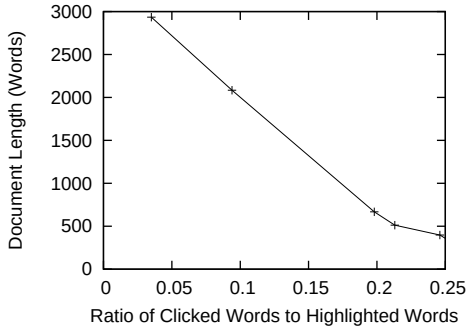

Figure 1. Document Length Affects Clicks

It should be noted that there is a large class imbalance in the data. For every click in document four, there are about 30 non-clicks. The situation is even more imbalanced for the science documents. For the second science document there are 100 non-clicks for every click and for the first science document there are nearly 300 non-clicks for every click.

There was also no correlation seen between a word being on a quiz and being clicked. This indicates that the students may not have used the system as seriously as possible and introduced noise into the click data. This is further evidenced by the quizzes, which show that only about 10% of the quiz words that students got wrong on the first test were actually learned. However, we will show that we are able to predict clicks regardless.

Figure 2, 3 and 4 show the relationship between the mean age of acquisition of the words clicked on, STAR language scores and the number of clicks for document 2. A second-degree polynomial was fit to the data for each figure. Students with STAR language scores above 300 are considered to have basic ability, above 350 are proficient and above 400 are advanced. Age of acquisition scores are abstract and a score of 300 means a word was acquired at 4-6, 400 is 6-8 and 500 is 8-10 (Cortese and Fugett, 2004).

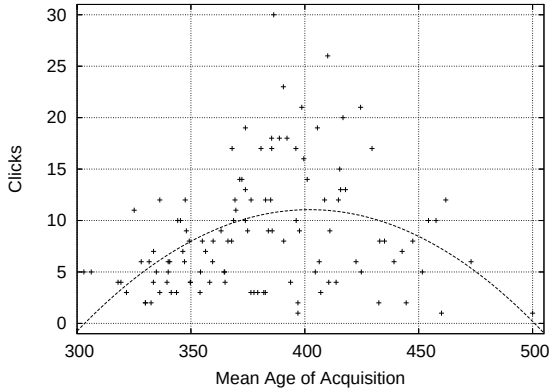

Figure 2. Age of Acquisition vs Clicks

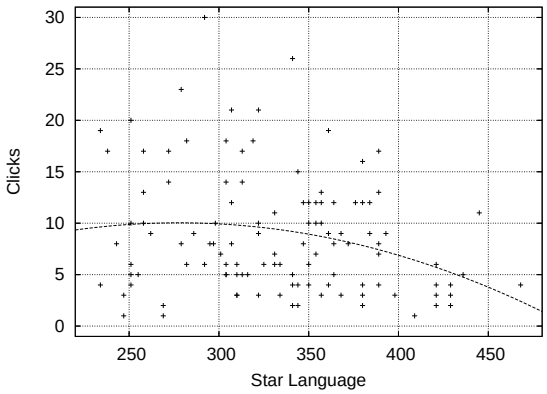

Figure 3. STAR Language vs Clicks

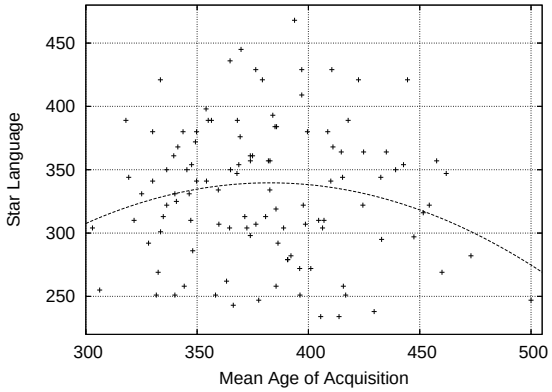

Figure 4. Age of Acquisition vs STAR Language

4 Machine Learning Method

The goal of our study is to predict student clicks in a vocabulary learning system. We used the random forest machine learning method, due to its success in the Yahoo! Learning to Rank Challenge (Chapelle and Chang, 2011). This algorithm was tested using the Weka (Hall et al., 2009) machine learning software with the default settings.

Random forest is an algorithm that classifies data by decision trees voting on a classification (Breiman, 2001). The forest chooses the class with the most votes. Each tree in the forest is trained by first sampling a subset of the data, chosen randomly with replacement, and then removing a large number of features. The number of samples chosen is the same number as in the original dataset, which usually results in about one-third of the original dataset left out of the training set. The tree is unpruned. Random forest has the advantage that it does not overfit the data.

To implement this algorithm on our click data, we constructed feature vectors consisting of both student features and word features. Each word is either clicked or not clicked, so we were able to use a binary classifier.

5 Evaluation

5.1 Features

To run our machine learning algorithms, we needed features for them. The features used are of two types: student features and word features. The student features we used in our experiment were the STAR (Standardized Testing and Reporting, a California standardized test) language score and the CELDT (California English Language Development Test) overall score, which correlated highly with each other. There was a correlation of about -0.1 between the STAR language score and total clicks across all the documents. Also available were the STAR math score, CELDT reading, writing, speaking and listening scores, grade level and primary language. These did not improve results and were not included in the experiment.

We used and tested many word features, which were discovered to be more important than the student features. First, we used the part-of-speech as a feature which was useful since only nouns were highlighted in the study. The part-of-speech tagger we used was the Stanford Log-linear Part-of-Speech Tagger (Toutanova et al., 2003). Second, various psycholinguistic variables were obtained from five studies (Wilson, 1988; Bird et al., 2001; Cortese and Fugett, 2004; Stadthagen-Gonzalez and Davis, 2006; Cortese and Khanna, 2008). The most useful was age of acquisition, which refers to "the age at which a word was learnt and has been proposed as a significant contributor to language and memory processes" (Stadthagen-Gonzalez and Davis, 2006). This was useful because it was available for the majority of words and is a good proxy for the difficulty of a word. Also useful was imageability, which is "the ease with which the word gives rise to a sensory mental image" (Bird et al., 2001). For example, these words are listed in decreasing order of imageability: beach, vest, dirt, plea, equanimity. Third, we obtained the Google unigram frequencies which were also a proxy for the difficulty of a word. Fourth, we calculated click percentages for words, students and words, words in a document and specific words in a document. While these features correlated very highly with clicks, we did not include these in our experiment. We instead would like to focus on words for which we do not have click data. Fifth, the word position, which indicates the position of the word in the document, was useful because position bias was seen in our data. Also important was the word instance, e.g. whether the word is the first, second, third, etc. time appearing in the document. After seeing a word three or four times, the clicks for that word dropped off dramatically.

There were also some other features that seemed interesting but ultimately proved not useful. We gathered etymological data, such as the language of origin and the date the word entered the English language; however these features did not help. We were also able to categorize the words using WordNet (Fellbaum, 1998), which can determine, for example, that a boat is an artifact and a lion is an animal. We tested for the categories of abstraction, artifact, living thing and animal but found no correlation between clicks and these categories.

5.2 Missing Values

Many features were not available for every word in the evaluation, such as age of acquisition. We could guess a value from available data, called imputation, or create separate models for each unique pattern of missing features, called reduced-feature models. We decided to create reduced feature models due to them being reported to consistently outperform imputation (Saar-Tsechansky and Provost, 2007).

5.3 Experimental Set-up

We ran our evaluation on document four, which had click data for 22 students. We chose this document because it had the highest correlation between a word being a quiz word and clicked, at 0.06, and the correlation between the age of acquisition of a word and that word being a quiz word is high, at 0.58.

The algorithms were run with the following features: STAR language score, CELDT overall score, word position, word instance, document number, age of acquisition, imageability, Google frequency, stopword, and part-of-speech. We did not include the science text data as training data. The training data for a student consisted of his or her click data for the other fables and all the other students' click data for all the fables.

5.4 Results

From Figure 2 we see the performance of random forest. We obtained similar performance with the other documents except document one. We also note that we also used a bayesian network and multi-boosting in Weka and obtained similar performance to random forest.

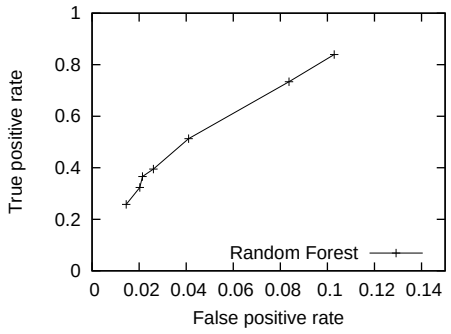

Figure 5. ROC Curve of Results

6 Discussion

There are several important issues to consider when interpreting these results. First, we are trying to maximize clicks when we should be trying to maximize learning. In the future we would like to identify which clicks are more important than others and incorporate that into our model. Second, across all documents of the study there was no correlation between a word being on the quiz and being clicked. We would like to obtain click data from users actively trying to learn and see how the results would be affected and we speculate that the position bias effect may be reduced in this case. Third, this study involved students who were using the system for the first time. How these results translate to long-term use of the program is unknown.

The science texts are a challenge for the classifiers for several reasons. First, due to the relationship between a document's length and the number of clicks, there are relatively few words clicked. Second, in the study most of the more difficult words were not highlighted. This actually produced a slight negative correlation between age of acquisition and whether the word is a quiz word or not, whereas for the fable documents there is a strong positive correlation between these two variables. It raises the question of how appropriate it is to include click data from a document with only one click out of 100 or 300 non-clicks into the training set for a document with one click out of 30 non-clicks. When the science documents were included in the training set for the fables, there was no difference in performance.

The correlation between the word position and clicks is about -0.1. This shows that position bias affects vocabulary systems as well as search engines and finding a good model to describe this is future work. The cascade model seems most appropriate, however the students tended to click in a non-linear order. It remains to be seen whether this non-linearity holds for other populations of users.

Previous work by Doğan and Lu in predicting click-words (Doğan and Lu, 2010) built a learning system to predict click-words for documents in the field of bioinformatics. They claim that "Our results show that a word's semantic type, location, POS, neighboring words and phrase information together could best determine if a word will be a click-word." They did report that if a word was in the title or abstract it was more likely to be a click-word, which is similar to our finding that a word at the beginning of the document is more likely to be clicked. However, it is not clear whether there is one underlying cause for both of these. Certain features such as neighboring words do not seem applicable to our usage in general, although it is something to be aware of for specialized domains. Their use of semantic types was interesting, though using WordNet we did not find any preference for certain classes of nouns being clicked over others.

Acknowledgements

I would like to thank Yi Zhang for mentoring and providing ideas. I would also like to thank Judith Scott, Kelly Stack, James Snook and other members of the TecWave project. I would also like to think the anonymous reviewers for their helpful comments. Part of this research is funded by National Science Foundation IIS-0713111 and the Institute of Education Science. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the author, and do not necessarily reflect those of the sponsors.

References

Helen Bird, Sue Franklin, and David Howard. 2001. *Age of Acquisition and Imageability Ratings for a Large Set of Words, Including Verbs and Function Words*. Behavior Research Methods, Instruments, & Computers, 33:73-79.

Leo Breiman. 2001. *Random Forests*. Machine Learning 45(1):5-32

Olivier Chapelle and Yi Chang. 2011. *Yahoo! Learning to Rank Challenge Overview*. JMLR: Workshop and Conference Proceedings 14 1-24.

Michael J. Cortese and April Fugett. 2004. *Imageability Ratings for 3,000 Monosyllabic Words*. Behavior Research Methods, Instruments, and Computers, 36:384-387.

Michael J. Cortese and Maya M. Khana. 2008. *Age of Acquisition Ratings for 3,000 Monosyllabic Words*. Behavior Research Methods, 40:791-794.

Nick Craswell, Onno Zoeter, Michael Taylor, Bill Ramsey. 2008. *An Experimental Comparison of Click Position-Bias Models*. First ACM International Conference on Web Search and Data Mining WSDM 2008.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman. 1990. *Indexing by Latent Semantic Analysis*. Journal of the American Society for Information Science, 41(6):391-407.

Rezarta I. Doğan and Zhiyong Lu. 2010. *Click-words: Learning to Predict Document Keywords from a User Perspective*. Bioinformatics, 26, 2767-2775.

Christine Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Yoav Freund and Robert E. Shapire. 1995. *A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting*. Journal of Computer and System Sciences, 55:119-139.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten 2009. *The WEKA Data Mining Software: An Update*. SIGKDD Explorations, Volume 11, Issue 1.

Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Geri Gay. 2005. *Accurately Interpreting Clickthrough Data as Implicit Feedback*. Proceedings of the ACM Conference on Research and Development on Information Retrieval (SIGIR), 2005.

Maytal Saar-Tsechansky and Foster Provost. 2007. *Handling Missing Values when Applying Classication Models*. The Journal of Machine Learning Research, 8:1625-1657.

Hans Stadthagen-Gonzalez and Colin J. Davis. 2006. *The Bristol Norms for Age of Acquisition, Imageability and Familiarity*. Behavior Research Methods, 38:598-605.

Kristina Toutanova, Dan Klein, Christopher Manning, Yoram Singer. 2003. *Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network*. Proceedings of HLT-NAACL 2003, 252-259.

Michael D. Wilson. 1988. *The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2*. Behavioural Research Methods, Instruments and Computers, 20(1):6-11.