

INFORMATION RETRIEVAL / TOPIC MODELING

Improving search in Wikipedia through quality and concept discovery

Christopher DeHoust · Praful Mangalath · Brian Mingus

University of Colorado Boulder

2008 technical report · CSCI 5417 course submission, December 23, 2007

Can the topics in an article help predict its quality and bring related Wikipedia articles into the same search results?

brianmingus.com

Research portfolio edition · The complete eight-page manuscript follows.

Conceptual cover artwork generated for this edition.

BRIAN MINGUS RESEARCH COLLECTION / BM-8FA40455A8F1

Improving search in Wikipedia through quality and concept discovery

Christopher DeHoust
Department of Computer Science
University of Colorado
Boulder, Co, 80309
DeHoust.Christopher@colorado.edu

Praful Mangalath
Department of Computer Science
University of Colorado
Boulder, Co, 80309
Praful.Mangalath@colorado.edu

Brian Mingus*
Department of Psychology
University of Colorado
Boulder, Co, 80309
Brian.Mingus@colorado.edu

Abstract

We investigate two complementary approaches to improving Wikipedia's search engine, each of which takes advantage of community efforts to ensure that the encyclopedia is tagged for quality and maintains a rich and highly coherent link structure. Using the Random Forest (RF) classifier [1], we attempt to predict community ratings of quality using surface-level features. The benefit of RF is that we can analyze the classifier's decision making process in order to determine the features underlying quality and derive a ranking based on these decisions. We use Latent Dirichlet Allocation (LDA) [2] to infer the topic structure underlying the encyclopedia and use those topics to not only predict quality, but to improve recall through pseudo-relevance feedback and clustering on top of the vector space model. While we are able to achieve modest performance in classification and extract interesting decision weights from RF, we find that the present rating scheme at Wikipedia lacks the distinctiveness necessary to use these quality heuristics in document ranking. We further find that the latent topics derived from LDA are just as good if not better at predicting community ratings, and that the same topics used in prediction can be used to improve a qualitative assessment of recall in search.

1 Introduction

The chaos of the Wikipedia editorial process and the broad coverage that it offers on a myriad of topics makes it a much debated subject, with one camp labelling it a 'free for all, a rumble tumble forum where interested people can get together in never-ending, circular conversations [3] and the other defending this randomness where the information is simply emergent and that a little slop at the microscale is the price of such efficiency at a macroscale [4].

Given this broad coverage of topics, some of which is good, most of which is bad, it is important that we are able to assess and rate articles before serving results. To this aim, we use the notion of concept learning, which involves looking within the documents and investigating how relevant the specific document is to the queried concept. We develop a document model based on the link

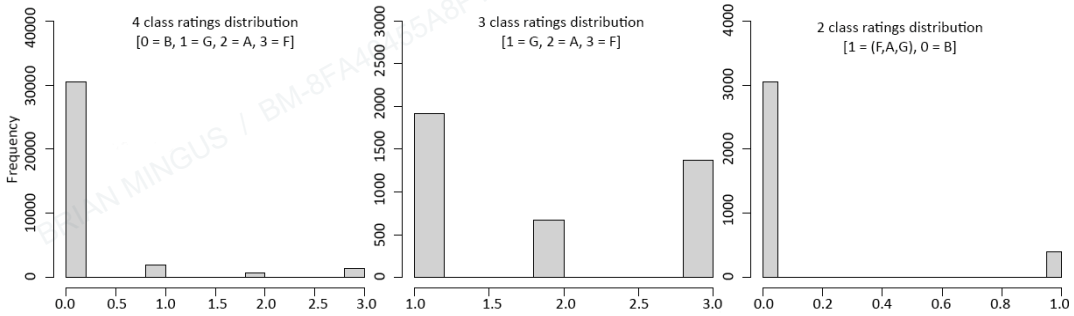
*All authors contributed equally to this work

BRIAN MINGUS RESEARCH COLLECTION / BM-8FA40455A8F1

structure, i.e., documents that are linked to and from a particular document and the text content of the document. Cluster based methods have proven to be an effective framework for information retrieval in the past few years [5]. They differ from simple vector space models by grouping documents into clusters and returning results based on this structure. We develop our document models based on the intuition that, if we can find good clusters and identify the highest quality documents in them, retrieval performance can be improved over the baseline vector space models as reported. We employ LDA, a generative model for text and links with an end representation similar in principle to cluster based retrieval. Lack of standard benchmarks on the Wikipedia dataset and prior work on using LDA in a non classification setting limits the scope of objectively capturing improvement in retrieval performance. We present a preliminary evaluation of the LDA model on the English Wikipedia and Simple English Wikipedia corpora.

2 Quality Assessment

The ratings dataset spans a million articles in the English Wikipedia corpus. The ratings are in increasing order of Stub, Start, B, Good, A, and Featured. Predicting Stubs and Starts is a trivial task and we are primarily concerned with being able to capture Featured (F), A (A), Good (G) and B articles. The distribution of each of the classes is shown below in Figure 0. As you can see, class B is a dominant class. We also test the model's ability to discriminate between classes, including i) F, A and G and ii) collapsed F,A,G vs B.



The task is to predict the quality of a new article given the observed articles and their ratings. One way to collect features that are representative of the quality of an article is to posit oneself as a rater. Surface level features that are likely to be important in determining an article's quality will need to cover both how readable the article is and how much content it has. Based on this intuition we computed multiple readability measures, including the Automated Readability Index, Gunning Fog Index, Flesch-Kincaid Grade Level, Flesch Reading Ease, Smog Grading and a few others. The correlation matrix of these readability metrics is listed in Figure 1.

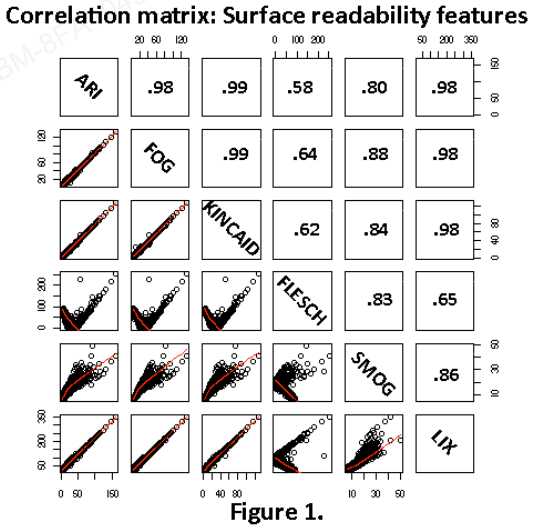


Figure 1.

BRIAN MINGUS RESEARCH COLLECTION / BM-8FA40455A8F1

The other surface level features we consider describe the quantity of text attributes, such as length, and counts of the number of paragraphs, words, references and sentences. We also used PageRank, computed from the Wikipedia article link matrix.The correlation matrix of these features is listed in Figure 2.

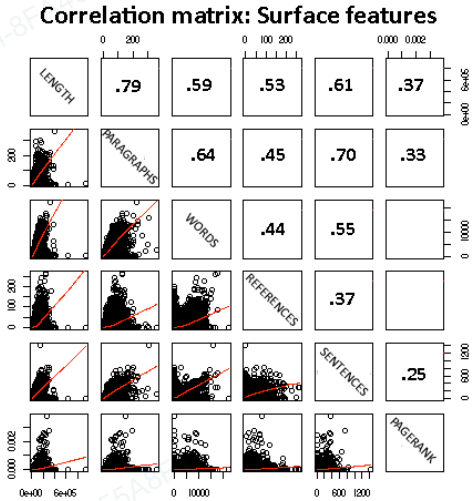


Figure 2.

As length is likely to be a good predictor, we also created a model based on the latent topic structure derived from LDA. The additional structure from the latent features will compensate for longer documents classified as B which have a dispersed representation or no structure. Instead of having a singular unit of content importance we now have a nice probabilistic interpretation of importance across the topics. The details of the LDA procedure and the dataset are described in Section 3.

2.1 Data Analysis

In this section we focus on the design of prediction models based on the features described. We use Support Vector Machines (SVM) [6] and Random Forests, two popular supervised learning methods for classification and regression. The goal of a SVM is to find a surface that maximally separates the data into articles that belong in a class, and articles that don't. Given a new representation of an article that is similar, but not identical to one of the classes that it was trained on, it is able to classify it based on its distance to the surface. RF's take an ensemble approach to learning by combining simple decision graphs in order to classify an article. Given a representation of an article, RF's grow hundreds of thousands of decision trees, each of which votes for a particular class. The model picks the class that has the most votes out of all the trees in the forest. RF's tend to be more resistant to inaccurate and noisy features as compared to individual trees, as the ensemble adjusts for the instability of individual trees that have been affected by small variations in the training set.

We performed SVM and RF classification on each of the 4, 3 and 2 class splits. We additionally reduced classification to a regression problem in order to compensate for the loss of continuity inherent in binary decisions. The heavily skewed distribution of the classes towards B makes it difficult to interpret the cross validation accuracy, so we also perform iterative sampling with 500 samples and 150 iterations after placing a uniform prior on each class. The intuition behind this is that the training and testing splits from 10-fold cross validation might not have enough coverage to test feature performance on each individual class. By forcing a uniform sample from each class we are able to define simple and naive performance benchmarks. The reported measures in the case of the entire dataset are:

- Regression (SVM & RF): Root Mean Square Error (RMSE) with 10 fold cross validation
- Classification (RF): Out of Bag (OOB) error rate
- Classification (SVM): Accuracy with 10 fold cross validation

BRIAN MINGUS RESEARCH COLLECTION / BM-8FA40455A8F1

The RMSE is the square root of the average of the squared differences between the actual and predicted class value. The class labels F,A,G,B direclyy translate to the scores 3,2,1,0. To interpret the RMSE values for a model, it is necessary to have a baseline system to compare against. This is done by assuming a naive system that would just look at all the articles and assign the average assigned rating from the training set. In the case of the RF's there is no separate cross validation procedure as an unbiased estimate is produced by leaving out a third of the cases during the construction of the decision trees. The prediction performance is the reported averaged measure on this hold out set, the OOB error rate. Accuracy for the SVM classification is the average percentage of correct matches of each run during cross validation. The iterative sampling process takes 500 random samples from each class for every iteration and computes the above measures. The reported data is the average across 150 iterations.

2.1.1 Four Quality Classes (F,A,G,B) Performance

For the 4 class case, a classifier which assigns B's to all articles will be 88% accurate, so a reasonable model should predict the class of a new article significantly above 88% accuracy. The best performance of a classifier was the RF on the LDA document count vectors at around 90%, dominated by the low class error on the B class shown in the confusion matrix. Analyzing this matrix makes it clear that the top three classes do not describe actual, predictable differences between documents.

		Correct Class				
		B	G	A	F	$ClassError$
Predicted	B	30362	70	4	89	0.005
	G	1663	132	0	122	0.931
	A	606	0	23	37	0.965
	F	985	60	0	329	0.760

The baseline RMSE of 0.67 in the case of the regression model is a naive approach which just predicts every rating as the mean of the entire training set. The RF classifier on the quality features had the lowest RMSE at 0.55, an improvement of 16.6% over the baseline.

In the uniform priors assumption case for the 4 class prediction, the baseline accuracy was 25% (error rate 75%) in the classification task, with the average RF on the quality features exhibiting the lowest error rate of 45.54%. Baseline RMSE in the regression case was 1.12 with the RF regression on the quality features showing a 19% improvement. Model performance for the 4 class task is listed below.

			Full Collection	**500 random samples from each class(150 iterations)**
Regression				
	Random Forest (RMSE)	LDA Document Vector (prob)	0.57	0.96
		LDA Document Vector (count)	0.57	0.97
		Quality Features	0.55	0.89
	SVM (RMSE)	LDA Document Vector (prob)	0.63	1
		LDA Document Vector (count)	0.63	1
		Quality Features	0.59	0.95
Classification				
	Random Forest (OOB error rate)	LDA Document Vector (prob)	10.1%	49.95%
		LDA Document Vector (count)	10.06%	54.96%
		Quality Features	10.57%	45.54%
	SVM (Accuracy)	LDA Document Vector (prob)	89.54%	42.54%
		LDA Document Vector (count)	89.54%	46.73%
		Quality Features	88.5%	50.35%

2.1.2 Three Quality Classes (F,A,G) Performance.

The three class baseline for the classification task on the entire dataset was an accuracy of 33.3%, with the quality features and RF classifier showing best performance. The RF quality features again showed best performance on the regression tasks with an improvement of 14.5% and 11.11% on the full and sampled dataset respectively. Model performance for the 3 class task is listed below.

		Correct Class			
		G	A	F	$ClassError$
Predicted	G	1486	36	395	0.22
	A	384	100	191	0.85
	F	427	16	931	0.32

BRIAN MINGUS RESEARCH COLLECTION / BM-8FA40455A8F1

			Full Collection	500 random samples from each class(150 iterations)
Regression				
	Random Forest (RMSE)	LDA Document Vector (prob)	0.81	0.74
		LDA Document Vector (count)	0.81	0.73
		Quality Features	0.77	0.72
	SVM (RMSE)	LDA Document Vector (prob)	0.86	0.78
		LDA Document Vector (count)	0.86	0.89
		Quality Features	0.82	0.76
Classification				
	Random Forest (OOB error rate)	LDA Document Vector (prob)	38.84%	44.22%
		LDA Document Vector (count)	38.84%	55.31%
		Quality Features	36.54%	40.95%
	SVM (Accuracy)	LDA Document Vector (prob)	55.16%	50.03%
		LDA Document Vector (count)	55.16%	57.33%
		Quality Features	60.79%	52.97%

2.2 Quality heuristics

The variable importance measures displayed below are taken from the Random Forest classifier on the 4 class full dataset. For each tree, prediction accuracy on the out of bag portion of the data is recorded. The same is done after permuting each predictor variable and the difference between the two accuracies are then averaged over all the trees and normalized by the total error.

The following table displays the relative contribution of each feature in predicting articles of either B, G, A or F. The final column shows the overall contribution made by the feature. It is important to note that these weights do not tell us the optimal value of the feature for that class. For example, although B class articles have a relatively high score of .04 on the Automated Readability Index, that does not mean that the higher an article scores on ARI, the more likely it is to be B. For all we know, it could be the opposite, or more likely, somewhere in-between.

Out of the 32 features assessed, readability measures dominate, with length related features coming in second. The ARI is the single best predictor of quality when averaged over classes. Featured articles are best described by a combination of the number of images, external and internal links, the Automated Readability Index and length, in that order. Interestingly, PageRank and age are poor predictors of quality overall.

	B	G	A	F	Overall Contribution
Automated Readability Index	0.04	0.01	0.00	0.02	0.04
Gunning Fog Index	0.03	0.00	0.00	0.01	0.03
Flesch-Kincaid Grade Level	0.04	0.01	0.00	0.01	0.03
Length of article HTML	0.03	-0.01	0.00	0.02	0.03
No. of paragraphs	0.02	0.00	0.00	0.00	0.02
Flesch Reading Ease	0.02	-0.01	0.00	0.00	0.02
Smog Grading	0.02	-0.01	0.00	0.00	0.02
No. of internal links	0.02	0.00	0.00	0.02	0.02
Laesbarhedsindex Readability Formula	0.03	0.00	0.00	0.01	0.02
No. of words	0.02	0.00	0.00	0.01	0.02
No. of references	0.02	0.01	0.01	0.02	0.02
No. of to be's	0.01	0.00	0.00	0.01	0.01
No. of sentences	0.01	0.00	0.00	0.01	0.01
Coleman-Liau Index	0.01	0.00	0.00	0.00	0.01
No. of templates	0.01	0.01	0.00	0.00	0.01
PageRank	0.01	0.00	0.01	0.00	0.01
No. of external links	0.01	0.02	0.01	0.04	0.01
No. of relative links	0.01	0.00	0.00	-0.01	0.01
No. of h2's	0.00	0.00	0.00	0.00	0.00
No. of h3's	0.01	0.00	0.00	0.00	0.00
No. of conjunctions	0.00	0.00	0.00	0.00	0.00
No. of images	0.00	0.01	0.01	0.06	0.00
Avg. word length	0.00	0.00	0.00	0.00	0.00
No. of h4's	0.00	0.00	0.00	0.00	0.00
No. of prepositions	0.00	0.00	0.00	0.00	0.00
No. of pronouns	0.00	0.00	0.00	0.00	0.00
No. of Interlanguage Links	0.01	0.00	0.00	0.00	0.00
Avg. syllables per word	0.00	0.00	0.00	0.00	0.00
No. of nominalizations	0.00	0.00	0.00	0.00	0.00
Article age (page-id)	0.00	0.00	0.01	0.00	0.00
Proportion of questions	0.00	0.00	0.00	0.00	0.00
Avg. sentence length	0.00	0.00	0.00	0.00	0.00

3 Concept assessment

In the standard tf-idf model employed by Lucene, documents are represented as a bag-of-words, where each document is a high dimensional vector over weighted words counts. [7] We approach the task of matching relevant documents to a query at the concept level by first deducing the topic

BRIAN MINGUS RESEARCH COLLECTION / BM-8FA40455A8F1

structure of the encyclopedia and then representing documents as probabilistic mixture over these topics. The goal is to improve recall even at a minimal cost to precision. Clustering is then used to capture finer distinctions than is possible with the sparseness of the vector space model alone. The motivation behind exploiting the latent features for document representation stems from the success LDA has had in TREC document classification tasks [8], and recently in Ad-Hoc retrieval [9] . We further extend the document model from the document word mixture setting into a document links mixture model.

3.1 LDA link model

The document-link model was experimental and only tested sufficiently on the Simple English Wikipedia corpus, a version of Wikipedia that consists of easy words and shorter sentences. The training data was a 22188 x 15254 sparse document-document matrix, with the (row,column) of the matrix set with value 1 if there is a link outbound from the document[row,:] to document[:,col] or vice versa. Applying LDA to this matrix of associations can be best described as a hierarchical clustering over link data where the latent variables help identify chunks of cohesive documents within the link network. Under this system the corpus can be visualized as a weighted network and the empirical strength of any document node can be estimated across the latent variables. The dimensionality was set to 400..

Latent Dirichlet Allocation (LDA) Google	Simple English Wikipedia LDA & Google & Simple	LDA & Google Google & Simple
Acid rock	Stanley Kubrick	Nuclear weapons
Aldeburgh Festival	1960s	Phish
Anne Murray	1966	Sexual liberation
Carl Radle	1967	Summer of Love
Harry Nilsson	Glam rock	Hippie
Jack Kerouac	South Park	Human Be-In
Phil Spector	African Americans	Students for a democratic society
Plastic Ono Band	Charles Manson	Woodstock Festival
Rock and Roll	Counterculture	Psychedelic Pop
Salvador Allende	Drug use	Lysergic acid diethylamide
Smothers Brothers	Flower Power	Summer of Love

The sample table above shows the results retrieved for the query “hippie” on

- The LDA link model. The 20 nearest neighbors of the document title matching the query.
- Google site search on simple.wikipedia.org. 20 results.
- The Simple English Wikipedia’s search engine. All outgoing and incoming links.

The above chart negates the hypothesis that the link model might be doing nothing more than capturing the direct local link structure already present in the incoming and outgoing hyperlinks on a page. It also shows the ability of the LDA model to capture associations which are not returned via a simple text match (from the lucene search) or from the google ’juice’. Similar tests were performed on several other keywords and the results returned followed the same trend. An official benchmarking was not possible due to the inherent bias in the subjects (authors) preferring results from the LDA model.

3.2 LDA text model

The corpus under consideration for the document-word model is the rated F,A,G,B English Wikipedia dataset containing 34491 documents and a vocabulary of 80953 words after stoplisting. The dimensionality was set to 800. Associations between document vectors and the query is taken as the conditional probability of the query given document, where the words in the query are treated as independent.

Sample results to the queries (’philosophy’,’economics’) are listed below for the LDA, Google site search and Wikipedia’s Lucene index. The Lucene and Google results tended to choose articles where the keywords were directly present, the intuition being a direct match is probably precise. While this works for longer queries, most users tend to query unigrams, opening up a big can of ’disambiguation’.

BRIAN MINGUS RESEARCH COLLECTION / BM-8FA40455A8F1

Latent Dirichlet Allocation (philosophy)	Lucene (philosophy)	Google (site:en.wikipedia.org philosophy)
Immanuel Kant	Philosophy	Philosophy
Critique of Pure Reason	Doctor of Philosophy	Greek philosophy
Philosophy	Political philosophy	Philosophy of mind
Aristotle	Greek philosophy	Chinese philosophy
Philosophy of Language	Analytic philosophy	Philosophy of science
Truth	Natural philosophy	Political philosophy
Empiricism	Continental philosophy	Metaphilosophy
David Hume	Philosophy of mind	Analytic philosophy
Ludwig Wittgenstein	Western philosophy	Philosophy of language
Pragmatism	Eastern philosophy	Philosophy of education
Georg Wilhelm Friedrich Hegel	Naturalism (philosophy)	Buddhist philosophy
Charles Peirce	Islamic philosophy	Islamic philosophy
Friedrich Daniel Ernst Schleiermacher	Philosophy of science	Eastern philosophy
Bertrand Russell	Chinese philosophy	Unix philosophy
Friedrich Nietzsche	Jewish philosophy	Western philosophy
Edmund Husserl	Philosophy of language	Renaissance philosophy
Philosophy of Friedrich Nietzsche	20th century philosophy	Hindu philosophy
Martin Heidegger	Philosophy of religion	Continental philosophy
Hilary Putnam	Philosophy of mathematics	Postmodern philosophy
Objectivism (Ayn Rand)	Iranian philosophy	Jewish philosophy

Latent Dirichlet Allocation (economics)	Lucene (economics)	Google (site:en.wikipedia.org economics)
Supply and Demand	London School of Economics	Category:Economics
Inflation	Ecological economics	Computational economics
Capitalism	Behavioral economics	Philosophy of economics
Competition Law	Labour (economics)	Ecological economics
Milton Friedman	Capital (economics)	Classical economics
United States Antitrust Law	Neoclassical economics	Keynesian economics
Monetary Policy	Welfare economics	Neoclassical economics
Minimum Wage	Keynesian economics	Welfare economics
Tax	Nobel Memorial Prize in Economic Sciences	Supply-side economics
Great Depression	Environmental economics	Financial economics
Monopoly	Agricultural economics	Development economics
Economy of Russia	Classical economics	Institutional economics
Fair Trade	Law and economics	Experimental economics
Economic Growth	Participatory economics	Economics
Heckscher-Ohlin Model	Marxian economics	Category:Economics
Money	Economic efficiency	Computational economics
Mercantilism	Consumption (economics)	Philosophy of economics
Business Cycle	Development economics	Ecological economics
Joseph E. Stiglitz	Financial economics	Classical economics

4 Conclusion

Using Random Forests allows one to add importance weights to quality heuristics, but the extent to which these weights are interpretable depends on the classes actually representing distinct differences between documents. In a previous study where five classes were used, [10] found that Start articles dominated. Our experiments, which use not only surface-level features, but also concept-level features, found that this trend continues in the four class task, where B dominates, and in the three class task, where G dominates. This trend stops at A articles because A is the least distinct class, not being very different from either F or G. The very features which make the lowest class articles distinct, such as being unreadable, short and not being a member of any link neighborhood, are no longer significant deficiencies in articles of greater classes and cease to be reliable predictors. Three features which the RF classifier considered to be reliable predictors of B articles, such as length, Flesh Reading Ease and Smog Grading, confuse the classifier on G articles. G articles are sometimes short, sometimes long, sometimes readable and sometimes unreadable. An article may be considered Good if it is unreadable as long as it makes up for it in length, or the situation could be reversed. Recognizing the non-linearity in such a task, we used the radial basis kernel in our SVM classifier, and it was still quite confused by these inconsistencies.

Despite this negative result, the success of using latent concepts to predict quality scores, which were just as good if not better than surface-level features, suggests that our approach works in principle, and would be successful given a more consistent rating system at Wikipedia. The ability of the concepts derived from the dataset to predict quality just as well as the surface-level features suggests that our classifiers perform as well as they can given the ratings. In the future, it may be possible to use LDA to not only complement the vector space model with clustering, but to select only the highest quality documents from that cluster for the return set.

BRIAN MINGUS RESEARCH COLLECTION / BM-8FA40455A8F1

References

[1] Leo Breiman. Random forests. *Machine Learning*, 2001.

[2] D. Blei; A. Ng; and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003.

[3] Let Wikipedia be Wikipedia. http://www.roughtype.com/archives/2005/. 2005.

[4] The Probabilistic Age. http://www.thelongtail.com/. 2005.

[5] Xiaoyong Liu and W. Bruce Croft. Cluster-based retrieval using language models. *Proceedings of SIGIR '04, Sheffield, England*, 2004.

[6] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm*, 2001.

[7] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. 1988.

[8] ChengXiang Zhai, Xinghua Lu, Xu Ling, Xin He, Atulya Velivelli, Xuanhui Wang, Hui Fang, and Azadeh Shakery. Uiuc/musc at trec 2005 genomics track. 2005.

[9] Xing Wei and W. Bruce Croft. Lda-based document models for ad-hoc retrieval. 2007.

[10] L. Rassbach; T. Blackford;B. Mingus;. Investigating the feasibility of automatically rating online article quality. *Technical Report*, 2007.

BRIAN MINGUS / BM-8FA40455A8F1

BRIAN MINGUS RESEARCH COLLECTION / BM-8FA40455A8F1