External Evaluation of Topic Models: A Graph Mining Approach

Hau Chan
Department of Computer Science
Stony Brook University
hauchan@cs.stonybrook.edu

Leman Akoglu
Department of Computer Science
Stony Brook University
leman@cs.stonybrook.edu

***Abstract*—Given a topic and its top-k most relevant words generated by a topic model, how can we tell whether it is a low-quality or a high-quality topic? Topic models provide a low-dimensional representation of large document corpora, and drive many important applications such as summarization, document segmentation, word-sense disambiguation, etc. Evaluation of topic models is an important issue; since low-quality topics potentially degrade the performance of these applications. In this paper, we develop a graph mining and machine learning approach for the external evaluation of topic models. Based on the graph-centric features we extract from the projection of topic words on the Wikipedia page-links graph, we learn models that can predict the human-perceived quality of topics (based on human judgments), and classify them as high or low quality. Experiments on four real-world corpora show that our approach boosts the prediction performance up to 30% over three baselines of various complexities, and demonstrate the generality of our method to diverse domains. In addition, we provide an interpretation of our models and outline the discriminating characteristics of topic quality.**

***Keywords*-topic models; human evaluation; graph mining**

I. Introduction

Topic modeling is an area that focuses on the extraction of topics from document corpora. Given a large collection of documents D and the number of desired topics T, a topic modeling method M, such as LDA [1], models each document $d \in D$ as a multinomial distribution over T topics, where each topic is in turn a multinomial distribution over W words. Typically, only a small number of words are important (i.e. have high likelihood) in each topic (also only a small number of topics are relevant for each document).

Topic models have been studied widely [1], [2], [3] and have applications in database summarization [4], word-sense discrimination [5], information discovery [6], etc. Naturally these applications rely on the quality of generated topics. An issue of concern, however, is that it is likely for topic models to output low-quality topics. For example, see Table I for two topics with their top 10 most likely words. From humans' perspective, the first topic (T1) consists of more semantically coherent words, while the second topic (T2) contains patchy groups of mostly incoherent words.

Low-quality topics can potentially degrade the performance of the applications; e.g. they could mislead topic-based document similarity, introduce noise in clustering, and cause poor semantic interpretation. This makes the evaluation of topic models a crucial task.

Previous research focused on the statistical (or quantitative) evaluation of topic models [7]. However these do not measure the interpretability of topics. In fact, [8] showed that there is a negative correlation between human vs. statistical evaluation of topic models. This finding started a new episode in topic model evaluation, by shifting the focus to semantic coherence of topics. Prior works on semantic evaluation of topic models include [8], [9], [10], [11]. None of these proposals (i) exploits a *collection* of evidential measures (they rather create a single measure), or (ii) builds a learning model to predict conceptual topic quality; which is the basis of our work. (See §IV related work details).

We introduce a graph mining approach for the external evaluation of topic models. We propose to use Wikipedia as an external resource, construct graph-centric features based on its page-links graph structure (referred to as *WikiLinks*), and build classification models to predict human-perceived quality of topics based on those evidential features. *WikiLinks* consists of articles about entities, which are linked by their relatedness, as perceived by human editors. Intuitively, we think of semantically coherent topics to consist of words that are "close-by" in this graph, and construct features based on graph topology and closeness accordingly. One key aspect of our framework is its generality; thanks to the domain-independent nature of our features. (see Figure 1)

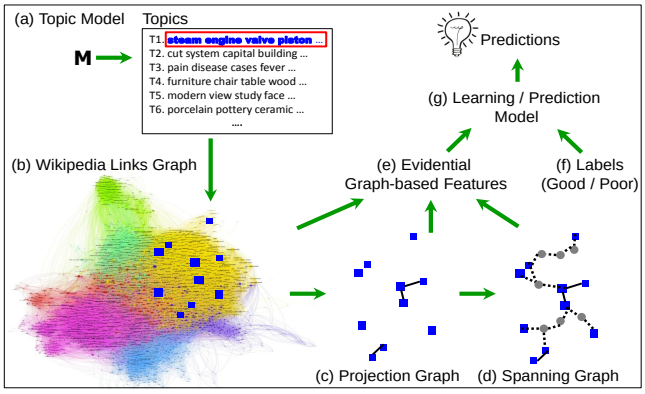


Figure 1. Proposed topic evaluation framework. (a) Given output topics by a topic model, (b) *WikiLinks* structure is leveraged to create (c) (induced) projection and (d) (connected) spanning graphs of topic words, then (e) graph-centric features are extracted for (f-g) learning predictive models.

Table I
EXAMPLE TOPICS T1 (HIGH-QUALITY) AND T2 (LOW-QUALITY) OF A TOPIC MODEL.

T1:	steam, engine, valve, piston, cylinder, pressure, boiler, air, pump, pipe
T2:	cut, system, capital, pointed, opening, building, character, round, france, paris

II. PROPOSED FRAMEWORK

A. Problem Definitions

We consider the topic quality prediction problem under two settings: (1) absolute and (2) relative quality prediction.

Our main problem aims to build models to predict the absolute, or the human-perceived, quality of the topics. Here, scores provided by several human judges determine the positive and negative class training labels.

- **(P1) Absolute (Human-Perceived) Quality Prediction:** Given a topic (i.e. a set of k words), predict its quality (good/poor) as judged by humans.

We also study a related classification task of predicting relative quality of topic words. Each topic output by a topic model consists of a sequence of K words sorted by their relevance to the topic. We treat the top-k (out of K) words of the topics as the positive (i.e. good) class examples, and bottom-k words as the negative (i.e. poor) class examples. Obtaining good prediction accuracy on this task would prove *WikiLinks* a good external resource.

- **(P2) Relative Quality Prediction:** Discriminate good versus poor quality topics defined by top-k versus bottom-k words, respectively.

B. Wikipedia Links Graph

Wikipedia page-links dataset contains internal links between Wikipedia articles (i.e. entities)[1]. As such, the page-links data lends itself for a graph representation (which we call the *WikiLinks* graph) in which nodes denote Wikipedia entities, and edges capture the internal link relations among the Wikipedia articles

For example, let us consider the entity `steam`. The corresponding Wiki-page can be found at http://en.wikipedia.org/wiki/Steam. Other Wiki-pages can be reached from this page by following hyperlinks on this page, e.g., the page on `piston` (http://en.wikipedia.org/wiki/Piston) and `mist` (http://en.wikipedia.org/wiki/Mist) are among those other, related entities. As such, the nodes `piston` and `mist` are 1-hop away from `steam`, thus are its direct neighbors.

WikiLinks is an excellent resource to guide for human-perceived evaluation of topic qualities, exactly because it is created by humans themselves—the entities are linked by their relatedness, as perceived by human editors.

Our key insight is to exploit the "graph-closeness" of related entities in *WikiLinks* to quantify the semantic quality of topics. Intuitively, the words of a semantically coherent topic, such as {`steam engine valve piston ...`}, would have high proximity in the *WikiLinks*. In fact, the wiki-page for `engine` directly links to `steam`, and `steam` links to `engine` through `steam-engine`, putting these two words respectively 1-2 hops away each direction.

[1]http://wiki.dbpedia.org/Downloads38#wikipedia-pagelinks

C. Projection and Spanning Graphs

We next provide definitions for topic subgraphs. Consider the *WikiLinks* graph $G(N, E)$ with node set N, edge set E (undirected or directed). Let W denote the set of k topic words, i.e. $|W| = k$. We project the topic words onto *WikiLinks* by *mapping* each word to a node (or entity) in the graph. In general, not all words will exist in *WikiLinks*, that is, $|N \cap W| \leq k$. We denote the mapped word set as $M = N \cap W \subseteq W$.

- **Topic projection graph** is a subgraph $g_M(M, E_M)$ induced on G with node set M and edge set E_M : $\{(u, v) \in E, u \in M \wedge v \in M\}$.

This graph may potentially consist of multiple disconnected components. In order to obtain a connected graph, we use a set of additional, connector nodes $C \subseteq N$ to build a graph that *spans* the topic words.

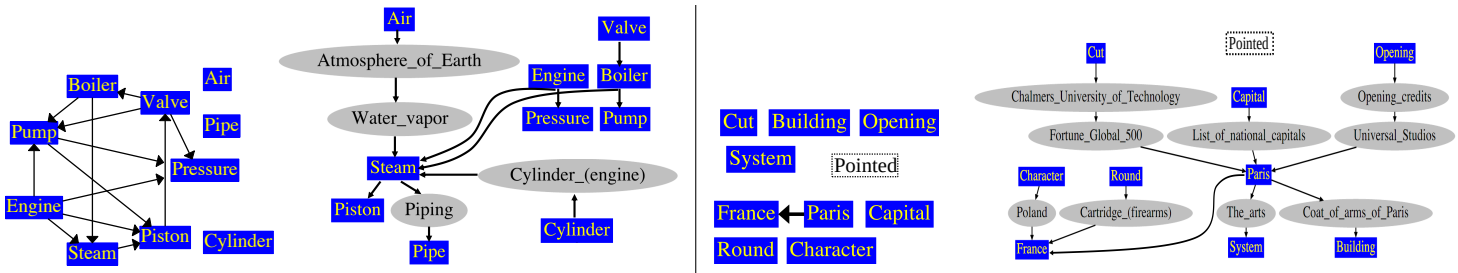
- **Topic spanning graph** is a subgraph $g_S(M \cup C, E_S)$ with node set $U = M \cup C$ and edge set $E_S : \{(u, v) \in E, u \in U \wedge v \in U\}$.

Ideally, the spanning graph contains the minimal set C to make the projection graph connected. However, it is NP-hard to find the minimal set, by reduction from the Steiner tree problem [12]). Therefore, we use the Minimum Spanning Tree (MST) approximation of the Steiner tree problem.

To construct the spanning graph, we first compute the pairwise shortest paths among $u, v \in M$ nodes to build a graph g_{SP} with edge weights (or shortest path lengths) $w(u, v)$. For undirected *WikiLinks*, g_{SP} is a complete graph as all nodes have paths from one to another (i.e. *WikiLinks* is a weakly connected graph). For directed *WikiLinks*, g_{SP} may contain missing edges as not all nodes have a *directed* path to others (i.e. *WikiLinks* contains multiple strongly connected components). Next we find the MST of g_{SP}, and *expand* each shortest path to include the (set of) connector nodes C. Note that the spanning graph may no longer be a tree but may contain loops due to the intersection of the connector node sets of the paths. In Figure 2, we show the projection and spanning graphs for the topics T1 and T2 of Table I.

D. From Topic Subgraphs to Graph-Centric Features

There are many features one could extract from a given graph. We want features that could potentially help differentiate good topics from poor ones. Good-quality topic words are conjectured to lie "close-by" in *WikiLinks*, reachable with many short paths from one another. On the other hand,



(a1) g_M of T1 (b1) g_S of T1 (a2) g_M of T2 (b2) g_S of T2

Figure 2. Projection and spanning graphs, g_M and g_S respectively, for the two example topics T1 and T2 as given in Table I. Blue square: mapped topic word $\in M$, dotted white square: missing word from *WikiLinks*, gray oval: connector node $\in C$.

the words of a poor topic would be separated in the graph topology. We can observe that these insights hold for T1 (good) and T2 (poor) of Table I in Figure 2. Specifically, T1 contains more words that exist in *WikiLinks* (i.e. words that map to *WikiLinks* nodes), consists of fewer connected components in its projection subgraph (i.e. more nodes with direct connection), requires fewer connector nodes to build its spanning graph, and so on. Using these observations, we construct features based on graph topology and closeness.

Table II gives the list of features we constructed. In total, we have 19 features capturing the key topological properties of the projection and spanning subgraphs, as well as the closeness measures of the topic words in the original *WikiLinks* graph.[2] We group our features into three: PROJ; features of the projection graph, D-SPAN; topological features of the directed spanning graph g_S, and D-SP; features capturing the pairwise reachability between the topic words.

E. Generating Case Libraries

We used news articles, books, and medical documents as our corpora. Descriptions of the datasets are in Table III.

For the prediction of the human-perceived (absolute) quality of topics (P1), we used the BOOKS and NEWS corpora, as previously used in [10], [11].[3] They consist of $T = 120$ and $T = 117$ topics, respectively. We considered the topics to consist of their top-10 words. All 237 topics were presented to 9 human judges. The judges were given guidelines on how to judge the *goodness* of the topics, and decide to what extent the topics were coherent, interpretable, meaningful, and easy-to-label with a short subject heading. They were also shown examples of good and bad topics. These nine judges evaluated the topics and provided annotations for each topic in 3-point scale: 1: 'good', 2: 'mediocre', 3: 'poor'. Topics with average rating below 1.5 are assigned to class 1 (good), and 0 (poor) otherwise. Examples of training topics from BOOKS (top few words) are given below (average rating in parentheses):

[2]Our experiments with directed and undirected versions of *WikiLinks* revealed that directed features provide more predictive power than undirected ones. Thus we focus our discussion on the directed features.

[3]We thank David Newman and his group for sharing the NEWS and BOOKS datasets as well as their human topic annotations.

Table II
EVIDENTIAL FEATURES USED IN MODEL LEARNING: PROJ, D-SPAN, AND D-SP EXTRACTED FROM PROJECTION GRAPH, SPANNING GRAPH, AND PAIRWISE SHORTEST PATHS OF TOPIC WORDS ON *WikiLinks*, RESPECTIVELY. (MST: MINIMUM SPANNING TREE)

***WikiLinks* Feature**	**Description**
PROJ: Topic projection graph (g_M) features (4)	
$g_M NumMiss$	missing words: $k - \|M\|$
$g_M NumConnComp$	connected components in g_M
$g_M SizeMaxComp$	nodes in largest component of g_M
$g_M MaxDeg$	maximum degree in g_M
D-SPAN: (Directed) Topic spanning graph (g_S) features (6)	
$g_S AvgMSTWeight$	average MST $W_{MST}/\|M\|$
$g_S RatioC$	ratio of $\|C\|/\|M\|$
$g_S MaxDegreeM$	maximum degree of M in g_S
$g_S MaxDegreeC$	maximum degree of C in g_S
$g_S AvgDegree$	average degree of nodes in g_S
$g_S Density$	density $\frac{\|E_S\|}{(\|M \cup C\|(\|M \cup C\|-1)))}$
D-SP: (Directed) Shortest path features (9)	
$NumNoPath$	num of pairs no DSP
$AvgSPLen$	average pairwise DSP length
$MaxSPLen$	maximum pairwise DSP length
$NumSP1$	num of pairwise DSP of length 1
$NumSP2$	num of pairwise DSP of length 2
$NumSP3$	num of pairwise DSP of length 3
$NumSP4$	num of pairwise DSP of length 4
$NumSP5$	num of pairwise DSP of length 5
$NumSP6+$	num of pairwise DSP of length ≥ 6

\+ silk lace embroidery tapestry gold embroidered ... (1)
\+ garden plant soil planting seed bloom spring ... (1.11)
\+ seed trees soil root planting plant tree ... (1.33)
– world people soul mind read reading live ... (2.56)
– white munich phil room student people head ... (2.67)
– person occasion purpose respect answer short ... (3)

We performed two measurements to quantify the inter-annotator agreement among the nine judges: (1) Average pairwise Spearman's rank correlation coefficient is found as $\rho = .73$ for NEWS and $\rho = .78$ for BOOKS; and (2) Average pairwise Cohen's kappa is found as $\kappa = .64$ for NEWS (max $\kappa = .79$), and $\kappa = .69$ for BOOKS (max $\kappa = .85$). While there is no precise rule for interpreting kappa scores, [13] suggests that scores in the range $(.60, .80]$ correspond to "substantial agreement" among the annotators.

For the prediction of the relative quality of topics (P2), we used the publicly available PRESS and BRAIN corpora and

Table III
DATASETS USED IN OUR EXPERIMENTS. D: # DOCUMENTS IN THE CORPUS, T: # TOPICS, $Labels$: WHETHER HUMAN ANNOTATIONS EXIST OR NOT.

Dataset	D	T	$Labels$	Description
BOOKS	12,000	120	Yes	Books downloaded from the Internet Archive
NEWS	55,000	117	Yes	NYTimes news articles from LDC Gigaword
PRESS	2,246	100	No	Documents from the Associated Press
BRAIN	10,000	200	No	Pubmed abstracts for the query "brain injury"

learned LDA [1] topic models with $T = 100$ and $T = 200$ topics, respectively. We considered the top-10 words for each topic to be in the positive (good) class. For the negative class, we built three case libraries with words of ranks [11-20], [31-40], and [91-100]. This way we constructed three different learning tasks each with 200 and 400 training examples for PRESS[4] and BRAIN[5], respectively. Examples of training topics from PRESS (top few words) are given below ([top 1-10] vs. [91-100]):

+ space soviet shuttle nasa launch mission earth venus ...
− jupiter day help report released days data laboratory ...
+ research scientists researchers animals project state ...
− defense usda caused two temperatures side agricultural ...
+ power cars heat oil fuel energy electricity day ...
− account total carbon year just united lower i plan ...

F. Learning to Predict

We train logistic regression classifiers with L_1 norm regularization. More specifically, we are given n training examples (n topics) $\{(x^{(i)}, y^{(i)}, i = 1, \ldots, n\}$, where each $x^{(i)} \in \mathbb{R}^m$ is an m dimensional feature vector, and $y^{(i)} \in \{1, 0\}$ denotes the class label (1: positive (good) vs. 0: negative (poor)). Logistic regression classifier models the probability distribution of the class label y given a feature vector x as $p(y = 0|x; w) = \sigma(w^T x) = \frac{1}{1+exp(-w^T x)}$, where $w \in \mathbb{R}^m$ are the parameters of the model (feature weights), and $\sigma(.)$ is the sigmoid function. We regularize the logistic regression model using L_1 norm, which corresponds to Bayesian learning under the Laplace prior of the parameters; $p(w) = (\lambda/2)^m exp(-\lambda \|w\|_1)$, with $\lambda > 0$. We report the leave-one-out cross-validation accuracies.

III. EXPERIMENT RESULTS

A. Relative Quality Prediction

The goal of this set of experiments is to understand the value of using the graph-centric evaluation framework we developed. Achieving promising performance on this pilot study would show us the feasibility of our approach. In Table IV, we present the prediction accuracy of our model on the PRESS and BRAIN topics. From the tables, we observe that using our graph-centric features we achieve improved classification performance in all cases, and when features are used collectively we obtain 15% to 30% boost over the random baseline. As expected, the boost is gradually higher for the easier tasks (from left to right) where the negative class words are chosen further down in the rank order of topic words. These preliminary results show that *WikiLinks* is useful as an external resource and that our method is suitable for topic quality prediction tasks.

B. Building baselines

Before we move on to results on human-perceived (absolute) topic quality prediction, we introduce and compare two non-trivial baselines to our approach.

1) Google baseline: Given a set of k topic words, we used several Google operators[6] to query for results containing these words. From four types of Google queries we built four what we call "Google features" per topic: (1) all the topic words in their text; allintext:$word_1, word_2, \ldots, word_k$ (2) at least one topic word in their title; intitle:$word_1$ OR ... OR intitle:$word_k$, (3) at least one topic word in their anchor; inanchor:$word_1$ OR ... OR inanchor:$word_k$, and (4) at least one topic word in their URL; inurl:$word_1$ OR ... OR inurl:$word_k$. For each topic word, we recorded the $\log(hitcount(query))$ for each of the Google features. Google features rely on an external resource; the Google search engine. However, they do not exploit graph-centric properties of any projection or spanning graphs.

2) PPR baseline: A second baseline classifier we built uses features based on the graph proximities among the topic words. To measure the proximity of a given pair of words on the *WikiLinks* graph, we used the personalized PageRank (PPR) scores [14]. Intuitively, the PPR score of a node v with respect to a given node u is high if there exist many, short paths between these two nodes. We constructed four PPR features capturing the pairwise graph-proximity between the topic words (excluding the self-pairs) as given in Table V. PPR-based features also exploit the underlying *WikiLinks* graph structure, and they are known as being more robust than shortest paths in capturing graph-centric proximities. However, PPR computations are expensive as they rely on the mixing of random walks with restarts on the input graph (in millions of nodes/edges). On the other hand,

Table V
PPR FEATURES GENERATED TO BUILD A BASELINE CLASSIFIER.

PPR Feature	Description
$AvgPPRscore$	average pairwise PPR score
$MedPPRscore$	median pairwise PPR score
$AvgPPRorder$	average pairwise PPR order
$MedPPRorder$	median pairwise PPR order

[4]http://www.cs.princeton.edu/~blei/lda-c/

[5]https://code.google.com/p/topic-modeling-tool/downloads/list

[6]http://www.googleguide.com/advanced_operators.html

Table IV
PRESS|BRAIN RELATIVE QUALITY PREDICTION RESULTS. CLASSIFICATION ACCURACIES FOR PREDICTING RELATIVE (TOP-k VERSUS NON TOP-k) TOPIC QUALITY, FOR VARIOUS GROUPS OF FEATURES.

Feature set **top-**10 **vs.**	**top-**[11-20]		**top-**[31-40]		**top-**[91-100]	
	PRESS	BRAIN	PRESS	BRAIN	PRESS	BRAIN
BASELINE-MAJORITY	0.500	0.500	0.500	0.500	0.500	0.500
PROJ	0.505	0.622	0.715	0.705	0.765	0.725
D-SPAN	0.650	0.687	0.760	0.740	0.805	0.762
D-SP	0.605	0.665	0.710	0.760	0.750	0.790
PROJ+D-SPAN	0.650	0.687	0.745	0.722	0.790	0.777
PROJ+D-SP	0.650	0.672	0.710	0.752	0.815	0.800
PROJ+D-SPAN+D-SP	0.660	0.687	0.735	0.752	0.810	0.807

Table VI
BOOKS AND NEWS ABSOLUTE QUALITY PREDICTION RESULTS. ACCURACIES FOR PREDICTING ABSOLUTE (HUMAN-PERCEIVED) TOPIC QUALITY, FOR VARIOUS GROUPS OF FEATURES.

Feature set	BOOKS	NEWS	BOOKS +NEWS
BASELINE-MAJORITY	0.610	0.521	0.549
BASELINE-GOOGLE	0.642	0.624	0.629
BASELINE-PPR	0.842	0.735	0.785
PROJ	0.875	0.812	0.848
D-SPAN	0.892	0.769	0.844
D-SP	0.883	0.786	0.852
PROJ+D-SPAN	0.883	0.795	0.844
PROJ+D-SP	0.892	0.795	0.848
PROJ+D-SPAN+D-SP	0.900	0.821	0.831

Table VII
CROSS-DOMAIN ABSOLUTE QUALITY PREDICTION RESULTS.

Train \ Test	BOOKS	NEWS
BOOKS	0.900	0.769
NEWS	0.867	0.821

computing our graph features is fast since projected graphs are fairly small, and finding the shortest paths takes only a few seconds as often times the mapped nodes are close-by and thus most of the graph need not be traversed. Therefore PPR is a strong but expensive baseline.

C. Absolute (Human-Perceived) Quality Prediction

In Table VI, we present our main results. We observe that *all* subsets of our feature groups outperform all three baselines. In particular, the Google baseline introduces 3-10% improvement in accuracy over the majority-class baseline, and the PPR baseline based on the *WikiLinks* graph structure yields up to 23% increase. While these demonstrate the value of *WikiLinks* for this task, PPR baseline is costly. On the other hand, all our graph-centric features introduce at least 25% and up to 30% boost over the majority baseline. In fact even the simplest group of our features PROJ, based on the immediate induced subgraph of topic words on the *WikiLinks*, outperforms the baselines alone. [7]

[7]Note that combined features do not always yield the best accuracy. We attribute this to the fact that learning with more features increases the size and complexity of our model space.

Table VIII
SELECTED FEATURES AND LEARNED COEFFICIENTS OF OUR L_1-REGULARIZED LOGISTIC REGRESSION MODEL FOR BOOKS AND NEWS. NEGATIVE (POSITIVE) COEFFICIENTS CONTRIBUTE TO THE ODDS OF A GIVEN TOPIC TO BE GOOD (POOR) QUALITY.

Selected Feature	Coef: BOOKS	Coef: NEWS
$g_M NumMiss$	0.0626	0.0918
$g_S RatioC$	0.2940	0.5909
$g_M MaxDeg$	−0.2921	−0.4541
$g_M SizeMaxComp$	−0.8667	− − −−
$g_S AvgMSTWeight$	− − −−	0.2598
$NumSP2$	−0.9685	− − −−

Cross-domain classification. In order to understand the generalization power of our framework, we learned a classification model using the BOOKS dataset and tested it on the NEWS dataset, similarly we also trained on NEWS and treated BOOKS as our test data. We show our results in Table VII. We observe that the cross-domain accuracies are fairly comparable to those of within-domain. This generalization power is particularly driven by our graph-centric features that are domain-independent.

Analysis of the prediction models. Finally, we study the characteristics of our learned models. As we use Lasso-regularization in our model training which lends itself to feature selection, we analyze the selected features (i.e. those with non-zero coefficients) for BOOKS and NEWS, as given in Table VIII. Features with positive coefficients contribute to the odds that a given topic is poor, whereas features with negative coefficients advocate for the topic being good. More specifically, we deduce that good topics are those with fewer missing mapped words onto *WikiLinks* (or larger M), fewer connector nodes C in their spanning graphs g_S, and higher degree nodes in their projection graphs.

IV. RELATED WORK

Topic modeling has been studied widely in machine learning [1], [15] (LDA, random projections), information retrieval [2], [3] (LSI, pLSA), and cognitive science [16].

Most works in quantitative evaluation of topic models [7] employ a variety of measures of model fit, such as estimating the likelihood of held-out documents or measuring the performance of an external task that is independent of the topic space such as information retrieval. While useful, these

methods ignore the evaluation of the interpretability and semantic meaning of the topics for users [8].

[10] propose a new measure called pairwise mutual information (PMI) of topic-words based on co-occurrence statistics of word-pairs in large external text corpora, and show that PMI scores of topics are highly correlated (Pearson's correlation) with human scores. [11] show that PMI outperforms a range of other topic-scoring measures such as those based on lexical similarity and similarity in a given ontology. In [17], the PMI model is extensively evaluated on various different genres and domains of corpora (news, books, National Institutes of Health (NIH) abstracts) and various external corpora (Wikipedia articles, Google 5-grams, pubmed.gov abstracts).

PMI-based evaluation, however, requires the entire scan of external documents to compute the co-occurrence count for every pair of topic-words which can be over 2 million Wikipedia articles and 1 trillion Google 5-grams. Moreover, the best correlation to human-perceived quality depends on the type of external corpora used (according to [17]: Google for books, Wikipedia for news, and pubmed.gov for NIH abstracts yield the best correlation).

[9] propose to use the original (i.e. training) corpus itself, which has been used for topic extraction, to compute a PMI-like score based on co-occurrence statistics of topic-words in the original document collection. This is interesting, as the reasons behind *not* using the training corpus in [10] was stated as "...instead of using the collection itself to measure word association..., we use a large external text data source to provide *regularization*". This, of course, comes with the same challenges as for PMI. [18] also used cohesion and specificity of the topics to define a conceptual topic relevance score based on a concept hierarchy (ontology).

Relevance-based [18] and PMI-like measures [10], [9] as well as others in [17] are all based on a *single* statistic, whereas we identify a collection of evidential features.

Finally, while not directly applicable to topic evaluation, related work include automatic topic *labeling* [19], [20], [21] where the goal is to find a single most representative phrase (i.e. topic label or name) for each topic. Most related work in data mining that has inspired our work is [22], which used graph mining for evaluating the quality of search engine results to user queries. Other related graph-based techniques include connection subgraphs [23], [24], [25] that aim to succinctly connect a subset of nodes in a given graph.

V. Conclusion

In this work, we introduced a graph based framework for the external evaluation of topic models. We proposed to use Wikipedia as an external resource to assess the semantic coherence of a given set of words. We constructed graph-centric features based on the closeness of topic words in the topology of Wikipedia's page-links graph structure, and derived predictive models that learn from human judgments to classify topics as good or poor quality as perceived by humans. Our results showed the effectiveness and generality of our approach in predicting and interpreting the human-perceived quality of topic models, where we achieved up to 30% better performance over three baseline predictors on four document corpora from diverse domains.

Acknowledgments

We thank the anonymous reviewers for their feedback. This material is based upon work supported by the Stony Brook University Office of the Vice President for Research and the National Science Foundation Graduate Research Fellowship. Any findings and conclusions expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding parties.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
[2] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. of American Soc. for Info. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.
[3] T. Hofmann, "Probabilistic latent semantic indexing," in *SIGIR*, 1999.
[4] A. Haghighi and L. Vanderwende, "Exploring content models for multi-document summarization." in *NAACL*, 2009, pp. 362–370.
[5] S. Brody and M. Lapata, "Bayesian word sense induction." in *EACL*, 2009.
[6] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. L. Griffiths, "Probabilistic author-topic models for information discovery." in *KDD*, 2004.
[7] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. M. Mimno, "Evaluation methods for topic models." in *ICML*, 2009.
[8] J. Chang, J. L. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei, "Reading tea leaves: How humans interpret topic models." in *NIPS*, 2009, pp. 288–296.
[9] D. M. Mimno, H. M. Wallach, E. M. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models." in *EMNLP*, 2011, pp. 262–272.
[10] D. Newman, S. Karimi, and L. Cavedon, "External evaluation of topic models," in *Australasian Doc. Comp. Symp.*, 2009, pp. 11–18.
[11] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *ACL*, 2010, pp. 100–108.
[12] R. M. Karp, "Reducibility among combinatorial problems." in *Complexity of Computer Computations*, 1972, pp. 85–103.
[13] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, 1977.
[14] T. H. Haveliwala, "Topic-sensitive pagerank," in *WWW*, 2002.
[15] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: applications to image and text data." in *KDD*, 2001.
[16] T. L. Griffiths, M. Steyvers, and J. Tenenbaum, "Topics in semantic representation," *Psychological Review*, 2007.
[17] D. Newman, Y. Noh, E. M. Talley, S. Karimi, and T. Baldwin, "Evaluating topic models for digital libraries." in *JCDL*, 2010.
[18] C. C. Musat, J. Velcin, S. Trausan-Matu, and M.-A. Rizoiu, "Improving topic evaluation using conceptual knowledge." in *IJCAI*, 2011.
[19] Q. Mei, X. Shen, and C. Zhai, "Automatic labeling of multinomial topic models." in *KDD*, 2007, pp. 490–499.
[20] J. H. Lau, K. Grieser, D. Newman, and T. Baldwin, "Automatic labelling of topic models." in *ACL*, 2011, pp. 1536–1545.
[21] I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene, "Unsupervised graph-based topic labelling using dbpedia." in *WSDM*, 2013.
[22] J. Leskovec, S. T. Dumais, and E. Horvitz, "Web projections: learning from contextual subgraphs of the web." in *WWW*, 2007.
[23] L. Akoglu, J. Vreeken, H. Tong, D. H. Chau, N. Tatti, and C. Faloutsos, "Mining connection pathways for marked nodes in large graphs," in *SIAM SDM*, 2013.
[24] C. Faloutsos, K. S. McCurley, and A. Tomkins, "Fast discovery of connection subgraphs," in *KDD*, 2004, pp. 118–127.
[25] H. Tong and C. Faloutsos, "Center-piece subgraphs: problem definition and fast solutions," in *KDD*, 2006, pp. 404–413.