Turning Down the Noise in Classrooms

Mariheida Córdova Sánchez *
Department of Computer Science
Purdue University
West Lafayette, IN.
cordovas@purdue.edu

Pinar Yanardag *
Department of Computer Science
Purdue University
West Lafayette, IN.
ypinar@purdue.edu

ABSTRACT

The use of micro-blogging in classrooms is a recently trending concept in computer-aided education. Micro-blogs offer an effective way of communication in large classrooms, and engage students in meaningful discussions. However, existing micro-blogging systems in education setting suffer from a few drawbacks. First, relevant content might be overwhelmed by irrelevant posts to the lecture which could jeopardize effective learning. Second, students might generate redundant content by posting similar questions to each other and create substantial information overload. Third, posts covering different aspects of the class might be left undiscovered due to real-time characteristics of micro-blogs. To address these issues, we present a principled approach for picking a set of posts that promotes relevant and diverse content while effectively turning down the noise created by redundant posts. We formulate this task as a submodular optimization problem for which we provide an efficient and near-optimal solution. We evaluate our framework on real micro-blog based classroom datasets and our empirical results demonstrate that our framework is effectively able to cover the most important and diverse content that is being discussed in classrooms.

1. INTRODUCTION

The use of micro-blogging in a classroom setting can be useful at engaging students as it provides an effective way of communication, especially in large classes. Using provided tools, students are able to share their questions or comments in real time during the class as well as helping each other by addressing the questions of fellow students. Thus, the use of micro-blogs not only deepens the understanding of the students, but also engages them to the class in real time.

Even though micro-blogging holds great promise, there are a few of drawbacks of using micro-blogging in a classroom setting. Due to the real time characteristics of micro-blogs, many students might be posting similar questions at the same time, and overload the timeline. This is an undesired behavior since instructors have limited time to investigate and answer questions, and searching for different questions to answer would be time consuming. Moreover, some questions might be irrelevant to the lecture content which could jeopardize effective learning. For instance, some students might ask questions about logistics of the class such as homework deadlines or exam dates, or simply post humorous comments that are intended to entertain fellow students. An ideal framework should penalize such irrelevant questions since they might distract students and cause them to lose focus during the class. Finally, students might have questions about different aspects of the lecture. Given that micro-blogs have limited amount of space for display, it is critical to cover a diverse range of topics.

In this paper, we propose a novel framework that ranks posted questions by maximizing the *relevancy* and *diversity*. We formulate our framework as a submodular function that exploits the property of *diminishing returns*. Our algorithm focuses on four key features:

- **Relevance**: We want to recommend the most relevant questions that are aligned with the lecture material.
- **Coverage and Diversity**: We would like to cover different aspects of the lecture by keeping the recommended set of posts as diverse as possible.
- **Efficiency**: Due to the inherent real-time behavior of micro-blogging environments, we would like our framework to be fast and efficient.
- **Optimality**: We would like to provide optimality guarantees for our solution. In particular, we would like our framework to give a solution that is at most $1 - 1/e$ away from the optimal solution.

2. RELATED WORK

Micro-blogging has been increasingly used as a tool for communication between students and instructors in classrooms over the recent years [4, 7, 10, 26]. Some of these works analyze micro-blogging in the context of language learning [4]. Twitter, as one of the most popular micro-blogging service prevalent these days, has also been exten-

* Both authors contributed equally.

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author's site if the Material is used in electronic media.
WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2891073

Figure 1: Word cloud generated from the lecture material (left) which focuses on keywords like `stock, company, growth, price`. Word cloud generated from the posts of students (right) which focuses on topics such as `stock`, `company` as well as some diverse topics like `nyse` and `nasdaq`. Size of the text is positively correlated to frequency of each item.

sively studied in a variety of applications including educational environments [7, 10] and topic modeling [28]. In addition, assessing the credibility of tweets has been investigated [17]. One of the key problems of using micro-blogs in a classroom setting is the overwhelmingly large number of responses from the students for a single question posted by the instructor. There have been a few approaches addressing this issue such as [6], which uses the correlation between questions and responses to identify the most relevant content. [5] proposes a text categorization approach that identifies relevant content by using multiple features associated with posts such as textual information and votes of the students. The problem of finding related questions given a particular question is investigated by several works including [8, 13, 14]. [27] proposes a syntactic tree matching approach to identify similar questions in community-based question-answering services such as Yahoo! Answers. Methods such as [25] considers the quality of answers derived from the expertise of participating users in order to select relevant content. [19] evaluates various scoring measures and proposes a composite technique that uses a dynamic query-specific framework. [23] takes the interests of the users into account in order to identify and recommend most relevant questions. However, none of these approaches focus on selecting non-redundant and diverse content in micro-blogs. Several submodular frameworks have been proposed in the literature, including [1, 9, 21]. [1, 21] assume that there is a click model available in order to assess relevancy, which is an unrealistic assumption for many applications since implicit feedback is often inaccessible. [9] proposes to select a set of relevant and diverse blog posts in blogosphere. Micro-blog posts often consist of short questions or comments which bring several additional challenges when assessing relevancy.

3. DATASET

This study uses the data collected using a tool that is designed as a micro-blogging platform where students can post questions or comments in a real-time setting and interact with the instructor or fellow students during the lectures. The dataset consists of content that is posted during eight semesters of an undergraduate course titled *Finance* with a total of 20,000 posts.

In addition to obtaining micro-blogging posts, accompanying lecture materials that are prepared by the instructor(s) during the corresponding eight semesters for this course are also collected. Figure 1 shows two word clouds that are generated for a random session of this course. As can be seen from the figure, there is a certain overlap between the content that the instructor is focused on and the content that students emphasized on, such as `stock, company, prices`.

However, one can see that there is also some content that students are focused on even though they are not intensively covered in the lecture, such as `nasdaq` and `nyse`. Therefore, the use of a micro-blogging platform allows students not only to extend or enhance the material covered in the lecture, but also enables students to ask and learn additional details about the concepts by asking questions. An important problem of using micro-blogs in education setting is the irrelevant content generated by students. Table 1 lists some examples of relevant and irrelevant questions asked by students.

Table 1: Examples of relevant and irrelevant posts

Relevant Posts	**Irrelevant Posts**
Can you share some examples about disposition effect?	Will there be an extension for homework 3?
What is the prospect theory?	Is today's class canceled?
How much money is usually invested by hedge funds?	Invest in pizza!
Why having no credit is considered bad?	Will this be on the midterm?

Therefore, an ideal framework should a) consider the correlation between the lecture material and the content generated by students, b) discard or give low priority to irrelevant content, and c) cover a diverse range of concepts that students are interested in.

4. METHODOLOGY

In this section, we describe the components that comprise our framework. First, we introduce the submodular framework that selects a set of most relevant and diverse content for a given lecture. Then, we discuss the relevancy

Table 2: An example to illustrate the notion of diminishing returns in our application where l represents the lecture material, and $p_1, \ldots, p_3$ represents the posts submitted by the students.

Content	Keywords
l	`stock` $\times$ 2, `market` $\times$ 1, `company` $\times$ 1, `insurance` $\times$ 3, `life` $\times$ 2
p_1	`stock` $\times$ 1, `market` $\times$ 1
p_2	`life` $\times$ 1, `insurance` $\times$ 1, `company` $\times$ 1
p_3	`stock` $\times$ 1, `company` $\times$ 1

component which expands certain keywords using semantic databases.

4.1 Submodularity

Submodularity is a discrete optimization method that shares similar characteristics with concavity, while resembling convexity. Submodularity appears in a wide range of application areas including social networks, viral marketing [15] and document summarization [16]. Submodular functions exhibit a natural *diminishing returns* property, *i.e.*, given two sets S and T, where $S \subseteq T \subseteq V \setminus v$, the incremental *value* of an item v decreases as the context in which v is considered grows from S to T.

More formally, submodularity is a property of set functions, *i.e.*, the class of functions $f : 2^V \rightarrow R$ that maps subsets $S \subseteq V$ to a value $f(S)$ where V is a finite ground set. The function f maps any given subset to a real number. The function f is called normalized if $f(\emptyset) = 0$, and it is monotone if $f(S) \leq f(T)$, whenever $S \subseteq T$. The function f is called submodular if the following equation holds for any $S, T \subseteq V$:

$$f(S \cup T) + f(S \cap T) \leq f(S) + f(T) \tag{1}$$

It has been shown that submodular function minimization can be solved in polynomial time [12], while submodular function maximization is an NP-complete optimization problem and intractable. However, it has been shown by [18] that the maximization of a monotone submodular function under a cardinality constraint can be solved near-optimally using a greedy algorithm. In submodular function maximization, we are interested in solving the following optimization problem:

$$A^* = \underset{A \subseteq V : |A| \leq k}{\operatorname{argmax}} \; f(A)$$

subject to a cardinality constraint k. If a function f is submodular, takes only non-negative values, and is monotone, then even though the maximization is still NP complete, we can use a greedy algorithm (see Algoritm 1) to approximate the optimum solution within a factor of $(1 - 1/e) \approx 0.63$ [18].

We formulate our task as a submodular optimization problem for which we also provide an efficient and near-optimal solution. First, let us motivate why we need to consider the notion of diminishing returns with an illustrative example. Table 2 lists an example lecture content, indicated by l and three example posts, p_1, p_2, and p_3. The content of each example is listed by the keywords it contains and the number

Algorithm 1 Greedy submodular function maximization with budget constraint

Require: V,k
Ensure: Selected set of posts S
1: Initialize $S \leftarrow \varnothing$
2: **while** $|S| \leq k$ **do**
3: $\quad v \leftarrow \operatorname{argmax}_{z \in V \setminus S}(f(S \cup \{z\}) - f(S))$
4: $\quad S \leftarrow S \cup \{v\}$
5: **end while**
6: **return** S

of times the keywords occur. Let us assume that the task we are interested in is to rank posts in a set $\mathcal{P} = \{p_1, p_2,$ and $p_3\}$. Let us define the gain as the number of terms a given post contains from the *uncovered* portion of the lecture material. One can see that selecting p_2 as the first post yields the highest gain, as it contains three uncovered terms from the lecture material, namely, *life, insurance, and company.* After selecting p_2, notice that the gain of selecting p_1 becomes 2 and that of p_3 becomes 1. This is simply because p_1 provides two uncovered terms while p_3 only provides one uncovered term, since the keyword *company* is already covered by p_2. In other words, due to the diminishing returns property, the gain of selecting post p_3 is reduced. Therefore, we select p_1 as the second post and p_3 as the third post. With this intuition in mind, we design a submodular framework that considers the relevancy of a given post to the lecture material as well as ensuring the diversity of the selected content.

Given a lecture l and a set of posts $\mathcal{P}$, the task we are interested in is to select k posts as *relevant* and as *diverse* as possible. Let us assume that we have a function f available to us, which simply takes a lecture l and a set of posts $\mathcal{P}$ and computes the total *gain* that set $\mathcal{P}$ represents. Then, given a new post v, one can compute the marginal gain of adding v into a set $\mathcal{P}$ by computing the difference between $f(l, \mathcal{P} + v) - f(l, \mathcal{P})$. In other words, one can compute the benefit of selecting v as the next post to cover. Next, we discuss the design of such an f function.

An ideal objective function should take a number of important aspects into account. First, it should promote posts that are relevant to the lecture material since we do not desire students to be distracted by irrelevant content. Second, it should consider the amount of contribution each post provides in order to push the most relevant posts to the top. Finally, it should encourage novelty by selecting the most diverse posts to cover. We define our submodular function $f(\mathcal{P}, l)$ as:

$$f(l, \mathcal{P}) = \sum_{p \in \mathcal{P}} \alpha_p \left(1 - e^{\sum_{i \in p} \beta_i}\right) \tag{2}$$

where $p \in \mathcal{P}$ represents an arbitrary post in $\mathcal{P}$, α_p represents the relevancy of post p to the lecture content, i represents an arbitrary feature in post p, and β_i represents the extend in which feature i is present in post p. This function can be seen as a special case of *probabilistic cover* proposed by [1] and holds submodular properties since it is monotonic, concave, and non-negative. Since our function f is submodular, the optimal solution can be found with a $1-1/e$ approximation guarantee by using a greedy algorithm [18]. The greedy algorithm simply starts with an empty set of S

Table 3: Examples of augmented keywords with DBpedia.

Keywords	Augmented keywords
hedge funds	`investment, hedge, investors, fund, trade, undertake, markets, regulators, assets, liquid, pension, foundations`
disposition effect	`disposition, sell, finance, investors, price, effect, shares, dropped, assets, anomaly, tendency`
prospect theory	`decisions, theory, prospect, gains, losses, model, heuristics, alternatives, outcomes, involve, economic, risk`
beardstown ladies	`beardstown, investment, club, ladies, inception, market, 1983, usa, illinois, stock, business, women, professional`
steve jobs	`apple, computer, revolution, executive, charismatic, american, pioneer, chairman, electronics, designer, inventor, influential, businessman`

and selects the post with the largest marginal gain. Then, in each iteration, the post that generates the maximum relative increase of the objective function is added to the selected post list S. In other words, in each iteration, the relative gain of adding each post $v \in \mathcal{P}$ to the set of selected posts S is recomputed and the post with the highest gain is selected. The algorithm terminates when a predefined budget k is reached.

Two important components of our submodular function are α_i and β_p values. Given a post p, β_i simply represents number of times feature i is present in post p. Given a lecture l and a post p, α_p is computed by taking the dot product between the feature vectors of the lecture, and the post. Next, we discuss how to create feature vectors for the lecture content and the posts.

4.2 Relevancy

Since we would like to promote posts that are most relevant to the course content, we compute the relevancy between each post and the lecture material. However, the length of the lecture material tend to be quite larger than the posts and directly computing a similarity between two texts would yield insufficient results. The intuition behind is that the post is probably only relevant to a portion of the lecture. Therefore, we follow a *sliding window* approach instead of using the entire lecture content as follows: given a window size w, we divide the lecture material into w-sized chunks. Then, we compute the similarity between each post and each chunk, and return the average similarity as the relevancy score.

A key component to compute a good relevancy score is to choose an appropriate feature representation. For this purpose, we experimented three traditional approaches, namely, bag of words, n-grams, and topic-modeling approaches.

A bag of words, or BOW, is a representation of text in which a sentence is represented as a set of words and their frequencies while disregarding the word order. n-grams are n number of terms that are used together. For instance, "credit card", "insurance companies" are two n-grams with $n = 2$. Topic modeling is a more sophisticated approach to represent text. We use Latent Dirichlet Allocation (LDA) [3] to find topics in posts and lectures. LDA is a generative model which assumes that each document $d \in D$ is associated with a K-dimensional topic distribution. In other words, each document d covers K latent topics, where each topic is defined as a distribution over words drawn from a Dirichlet distribution $\phi_k \sim Dirichlet(\beta)$. Elements of ϕ_k denotes the probability that a particular word is used for that topic. We use Online LDA [11], an improvement over LDA which uses variational inference instead of a Collapsed Gibbs Sampler[1] for practical purposes. After applying LDA, each lecture and post is represented as a probability distribution over K topics.

In order to further improve the relevancy of our feature representation, we augmented each post by using external resources, following a similar spirit to *query expansion*. The intuition behind our approach is that, given a keyword that appeared in a question, we can use an external source to augment this piece of information and obtain expanded keywords. We adopted DBpedia [2], which is a crowd-sourced database that extracts structured information from Wikipedia. For a given keyword, we queried DBpedia and obtained a list of possible Wikipedia articles. Each Wikipedia article is ranked by the number of in-links pointing from other Wikipedia articles. Table 3 illustrates a list of keywords that are expanded using DBpedia. As we can see from the table, given a keyword **beardstown ladies**, we are able to obtain the information that this term is also related to other keywords such as `investment, inception, market, illinois`.

5. EXPERIMENTS

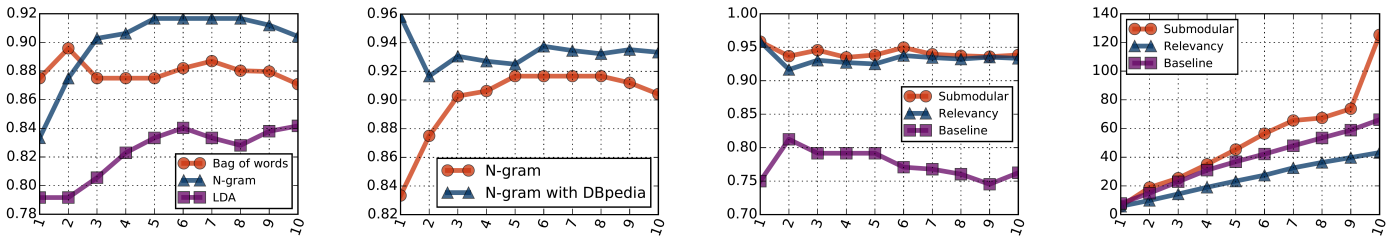
In this section, we first perform experiments to determine the best feature representation for our application. After that, we use the selected feature representation in our submodular function and compare the efficacy of our framework against other methods.

5.1 Feature representation

In order to determine which feature representation is more appropriate for our application, we compared *Precision@k* for BOW, n-grams and LDA methods. We used a window size of 200 for all methods, $n = 2$ for n-gram representation, and number of topics $K = 25$ for LDA.

Precision@k represents the fraction of the posts retrieved that are relevant to students at the top k results. In order to obtain a ground truth which is necessary to compute

[1]We used Gensim Python Library[22] for the model estimation process, also available at https://pypi.python.org/pypi/gensim



(a) (b) (c) (d)

Figure 2: Precision vs Rank using relevancy model for different feature representations (a). Precision@k results for feature augmentation experiment (b). Precision values at different ranks for model comparison (c). Diversity values at different ranks for model comparison (d).

Precision@k values, we manually labeled 1000 posts as relevant and irrelevant. Figure 2 (a) shows the comparison of *Precision@k* values for each method. As can be seen from the figure, the best results are obtained by using n-grams. We conclude that BOW representation falls short since it can not determine composite terms that are relevant to the lecture material such as `credit card`.

Moreover, we conclude that LDA does not perform well since applying LDA on short documents such as tweets, text messages or micro-blog questions is a challenging task. Previous efforts mainly focused on tweets where researchers applied methods including aggregating all the tweets of a user into a single document [24] which follows an author-topic model. However, this model fails to capture the fact that each tweet has its own topic assignment. Latest approaches such as Twitter-LDA [28] proposes to overcome this issue; however, it assumes that a single tweet is about a single topic and fails to capture the fact that a post can be about multiple topics. Labeled LDA [20] is another LDA-based approach, however, this model relies on labeled data such as hashtags which makes the model inherently inapplicable for our data since we do not have hashtags associated with posts. Therefore, we decided to use n-gram as our feature representation.

Figure 2 (b) shows the comparison between naïve n-gram approach and augmented n-gram approach. As can be seen from the figure, n-gram with DBpedia significantly improved *Precision@k* values since it uses an external source to augment the representation of individual words. Therefore, for the rest of our experiments, we use the feature representation with n-grams augmented by DBpedia.

5.2 Submodular framework

Our experimental setup is as follows. For each class setting that consists of a list of posts asked by students and a lecture material, we re-rank the questions using each of the following three methods:

- **Baseline**: Posts are left in their original position as they arrive to the system.
- **Relevancy**: Posts are sorted by their relevancy to the lecture content without considering diversity.
- **Submodular**: Posts are sorted by our submodular framework.

We computed *Precision@k* and *Diversity@k* values out of $k \in \{1, \ldots, 10\}$ for each class, and report the average value for eight available semesters. *Diversity@k* metric is computed by counting the number of unique terms at top k, which indicates the number of concepts covered by each method.

Figure 2 (c) shows the comparison of the methods with *Precision@k* values. As can be seen from the figure, the Submodular method outperforms the Relevancy method, while both Submodularity and Relevancy significantly outperform the Baseline method.

Figure 2 (d) shows the comparison of methods with *Diversity@k* values. As can be seen from the figure, Submodular method covers a significantly larger amount of unique terms compared to the Relevancy and Baseline methods. An interesting observation is that the Baseline method is more diverse than the Relevancy method. This is due to the fact that even though the Relevancy method selects posts that are relevant to the lecture material, it ends up selecting posts that are similar to each other and do not provide enough coverage of topics.

6. CONCLUSIONS

In this paper, we proposed a novel framework which finds the most relevant and diverse content in educational online discussions. We addressed main problems that occur in these scenarios: a) a large amount of irrelevant content, b) repeated or similar content, and c) content with not enough coverage or diversity. We proposed a submodular framework that ranks questions submitted by students by their relevance and diversity. Our empirical analysis shows the effectiveness of our proposed framework.

Moreover, our framework is not only applicable to micro-blogs in educational setting, but also to micro-blogs and question/answer websites in general.

As future work, we consider the profile of the students into account. In particular, one can rank questions based on how authoritative or active a student is in the class. In addition to enhancing the importance of the questions with student profiles, we also consider designing a *personalized* ranking for individual students to promote questions on topics that they are interested in.

References

[1] A. Ahmed, C. H. Teo, S. Vishwanathan, and A. Smola. Fair and balanced: Learning to present news stories. In *WSDM*, 2012.

[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. *Dbpedia: A nucleus for a web of open data*. Springer, 2007.

[3] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 2003.

[4] K. Borau, C. Ullrich, J. Feng, and R. Shen. Microblogging for language learning: Using twitter to train communicative and cultural competence. In *ICWL*, 2009.

[5] S. Cetintas, L. Si, H. Aagard, K. Bowen, and M. Cordova-Sanchez. Microblogging in a classroom: Classifying students' relevant and irrelevant questions in a microblogging-supported classroom. *TLT*, 2011.

[6] S. Cetintas, L. Si, S. Chakravarty, H. Aagard, and K. Bowen. Learning to identify students' relevant and irrelevant questions in a micro-blogging supported classroom. In *ITS*, 2010.

[7] C. Costa, G. Beham, W. Reinhardt, and M. Sillaots. Microblogging in technology enhanced learning: A use-case inspection of ppe. In *SIRTEL*, 2008.

[8] H. Duan, Y. Cao, C. yew Lin, and Y. Yu. Searching questions by identifying question topic and question focus. In *HLT*, 2008.

[9] K. El-Arini, G. Veda, D. Shahaf, and C. Guestrin. Turning down the noise in the blogosphere. In *SIGKDD*, 2009.

[10] G. Grosseck and C. Holotesku. Can we use twitter for educational activities? In *eLSE*, 2008.

[11] M. D. Hoffman, D. M. Blei, and F. Bach. Online learning for latent dirichlet allocation. In *NIPS*, 2010.

[12] S. Iwata, L. Fleischer, and S. Fujishige. A combinatorial strongly polynomial algorithm for minimizing submodular functions. *JACM*, 2001.

[13] J. Jeon, W. B. Croft, and J. H. Lee. Finding semantically similar questions based on their answers. In *SIGIR*, 2005.

[14] V. Jijkoun. Retrieving answers from frequently asked questions pages on the web. In *CIKM*, 2005.

[15] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *SIGKDD*, 2003.

[16] H. Lin and J. Bilmes. A class of submodular functions for document summarization. In *HLT*, 2011.

[17] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz. Tweeting is believing?: understanding microblog credibility perceptions. In *CSCW*, 2012.

[18] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions-I. *Mathematical Programming*, 1978.

[19] D. P. and S. Chakraborti. Finding relevant tweets. In *WAIM*. 2012.

[20] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*, 2009.

[21] K. Raman, P. Shivaswamy, and T. Joachims. Learning to diversify from implicit feedback. In *WSDM Workshop on Diversity in Document Retrieval*, 2012.

[22] R. Rehrek, P. Sojka, et al. Software framework for topic modeling with large corpora. University of Malta, 2010.

[23] Y.-I. Song, C.-Y. Lin, Y. Cao, and H.-C. Rim. Question utility: A novel static ranking of question search. In *AAAI*, 2008.

[24] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *SIGKDD*, 2004.

[25] M. A. Suryanto, E. P. Lim, A. Sun, and R. H. Chiang. Quality-aware collaborative question answering: methods and evaluation. In *WSDM*, 2009.

[26] C. Ullrich, K. Borau, H. Luo, X. Tan, L. Shen, and R. Shen. Why web 2.0 is good for learning and for research: Principles and prototypes. In *WWW*, 2008.

[27] K. Wang, Z. Ming, and T.-S. Chua. A syntactic tree matching approach to finding similar questions in community-based qa services. In *SIGIR*, 2009.

[28] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *ECIR*. 2011.