Thesis Proposal

Concept Modeling and Personalization for Combating Information Overload

Khalid El-Arini

November 19, 2010

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Carlos Guestrin, Chair
Zoubin Ghahramani
Tom Mitchell
Noah Smith
Thorsten Joachims, Cornell University

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Copyright © 2010 Khalid El-Arini

Abstract

Today, information overload is a ubiquitous problem that affects nearly all members of society, from researchers sifting through millions of scientific articles to Web users trying to gauge public opinion by reading blogs. While seemingly different, information retrieval tasks such as these often share characteristics that allow for common abstractions and solutions. In this thesis, we explore how insights from machine learning can shed light on three facets of such an abstraction: 1) representation, 2) modeling and optimization, and 3) interaction.

We represent a document collection as a set of *concepts* comprising the atomic units of information to be reasoned about. These concepts can range from low-level words, named entities and n-grams to higher level topics from a topic model. By assuming that each document in the collection covers a subset of concepts to some degree, we can model a variety of information retrieval tasks as *submodular coverage problems*, selecting a set of documents that maximally covers concepts of interest. While such coverage problems can be solved efficiently and near-optimally, they depend on a concept representation that models the ideas in a document collection at an appropriate level of granularity for each user. Leveraging Bayesian nonparametric methods and online learning, we intend to automatically learn such concept representations that are also robust to dynamism in the document collection.

Moreover, whether in terms of expressing an information need or providing feedback on results, proper user interaction is critical to the success of an information retrieval system. As users have different interests and trust preferences, we wish to provide personalized solutions by estimating how interested users are in particular concepts and how much these interests change over time, based on limited feedback.

We believe that by modeling document collections with an appropriate concept representation, we can effectively deal with information overload by framing efficient optimization problems that produce personalized solutions to common retrieval tasks with minimal burden on the user.

Contents

1 Introduction **1**
- 1.1 Representation . . . 2
- 1.2 Modeling and Optimization . . . 3
- 1.3 Interaction . . . 4
- 1.4 Thesis Statement and Research Questions . . . 6

2 Prior Contributions **8**
- 2.1 Turning Down the Noise in the Blogosphere (Q3, Q5) . . . 8
  - 2.1.1 Coverage (Q3) . . . 8
  - 2.1.2 Personalization (Q5) . . . 11
  - 2.1.3 Evaluation . . . 14
  - 2.1.4 Summary . . . 20
- 2.2 Beyond Keyword Search: Discovering Relevant Scientific Literature (Q3, Q4, Q6) . . . 20
  - 2.2.1 Expressing information needs (Q4) . . . 20
  - 2.2.2 Complex coverage modeling (Q3) . . . 21
  - 2.2.3 Trust modeling (Q6) . . . 21
  - 2.2.4 Evaluation . . . 23
  - 2.2.5 Summary . . . 25

3 Proposed Work **26**
- 3.1 Concept Granularity and Dynamics (Q1, Q2) . . . 26
  - 3.1.1 The Farmers Market Process . . . 27
  - 3.1.2 Redundancy Feedback and Personalized Concepts . . . 28
- 3.2 Transferred Feedback for Trust (Q6) . . . 29
  - 3.2.1 Twitter-based Trust . . . 29
- 3.3 Modeling Personalization Dynamics (Q5, Q7) . . . 30
  - 3.3.1 Changing User Interests . . . 30
  - 3.3.2 Overpersonalization . . . 31
- 3.4 Structured Input and Output (Q3, Q4) . . . 31
- 3.5 Expected Contributions . . . 32

4 Detailed Timeline **34**

Bibliography **35**

Chapter 1

Introduction

Today, many domains face the critical problem of information overload. Web users trying to gauge public opinion or learn about current events face a torrent of information from tens of millions of sources worldwide. For example, the blog indexing service Spinn3r retrieves over one million news articles, blog posts and social media updates every hour [1]. Similarly, for generations, scientists have built upon the published work of their predecessors and contemporaries in order to make new discoveries. However, today, with tens of millions of articles published in tens of thousands of journals and conferences [2], researchers face an acute difficulty in sifting through related literature.

As early as 1755, the French philosopher Denis Diderot presciently forewarned that there would come a day when "it will be almost as convenient to search for some bit of truth concealed in nature as it will be to find it hidden away in an immense multitude of bound volumes," [20]. Whether we take this "immense multitude" today as referring to corpora of online documents as described above, or other large multimedia collections on the Web, these information overload problems and their associated information retrieval tasks can be broken down into three primary facets:

1. *Representation.* How do we represent the information in a document collection to best facilitate a particular retrieval task?

2. *Modeling and Optimization.* How can the inherent structure of a particular retrieval task be incorporated into an efficiently optimized objective function, while at the same time providing theoretical guarantees on solution quality?

3. *Interaction.* How should a user's information need be specified? In which ways can a user provide feedback in order to indicate his or her interests and trust preferences in the least burdensome of manners?

In this thesis, we explore how insights from machine learning and optimization can allow us to expand the state of the art in information retrieval by addressing these facets in novel ways.

[1] `www.spinn3r.com`
[2] Statistics from Thomson Reuters: `http://wokinfo.com/about/whatitis/`

1.1 Representation

In order for a computer to reason about a document, a representation must be chosen for the document's contents. One of the simplest and most common representations for document modeling is known as the "bag of words" model, where the order of the words in a document is ignored, and the content of a document is simply represented as a list of words and their frequencies [31]. Some more sophisticated representations count n-grams, named entities, noun phrases and richer syntactic structures, while others represent a document as being generated from latent factors known as topics [5].

What these representations have in common is that they distill the contents of a document collection into fundamental *concepts* that represent the atomic units of information to be reasoned about. However, these concepts exist at imperfect granularities. In some cases, they may be too fine-grained, such as individual words or named entities, leading to many concepts with the same meaning. For instance, "President Obama," "Barack Obama," "Mr. Obama," "Mr. President" and "POTUS" are all distinct named entities that refer to the 44th President of the United States, but may end up as distinct concepts depending on the representation. On the other end of the spectrum, high-level concepts, such as topics from a probabilistic topic model, may confuse together many ideas that are only vaguely related.

As will become clear in the next two sections, appropriate concept granularity is important for information retrieval tasks due to the central role concepts play in both accurately modeling a retrieval problem and in understanding user feedback. For example, as diversity is often a desirable characteristic when selecting a set of documents to satisfy a user's information need, a retrieval algorithm must recognize when different documents talk about the same thing. Representing a document collection with concepts that are too fine-grained can result in redundant documents being selected in such a setting. Likewise, in information retrieval, we often wish to infer how interested a particular user might be in various concepts, so that we can select a set of personalized results. In this situation, measuring a user's preference for an overly broad concept makes little sense, as the user is unlikely to have a single, consistent opinion on such a vague collection of ideas.

In this thesis, we wish to *automatically learn* concept representations of document collections at levels of granularity appropriate for the information retrieval tasks at hand.

Additionally, different users may find different granularities of concepts more appropriate. To an avid sports fan, for example, the Pittsburgh Steelers, LeBron James and Wimbledon all represent distinct concepts from the world of sports. However, another user who may not care so much about sports may consider these all to be equivalent from his point of view. Namely, for this user, an article about the Steelers might just as well be an article about tennis; it's all "just sports" to him.

We wish to investigate the differences in optimal concept representations across users.

Moreover, document collections are more often than not dynamic, with new blog posts or research papers coming in by the hour or minute.

We wish to investigate how concepts change in real-world dynamic document collections.

Related Work There is a large body of work in the psychology literature about how people represent concepts [47]. For instance, we can view the approach we propose in Section 3.1.1 as related to early work in cognitive science on featural representations of categories (e.g., people might remember that bears often have the features "fur" and "claws") [29, 32]. As we expect that the appropriate level of

granularity for concepts lies somewhere in between words and topics from a topic model, we like to make the analogy with the super-pixels of the computer vision community. In some sense, we are trying to learn a representation for "super-words."

A large subfield of classical artificial intelligence is that of knowledge representation [9] and ontology learning [27]. While related in principle, these approaches tend to be tailored to more formal reasoning, e.g., for logic-based reasoning or automation. We seek to develop a more flexible and fluid representation of concepts. Likewise, there is work in the literature on rough sets [18] that attempts to learn hierarchies of concepts at various granularities, but these techniques have no probabilistic interpretation, and they assume documents are pre-classified into groups or clusters.

Dimensionality reduction techniques such as latent semantic analysis [17] or clustering-based methods [19] suffer from the same problem as the high-level topics from a topic model, in that the resulting dimensions or factors are hard to interpret.

To the best of our knowledge, little, if any, work has been done on personalizing concepts to individual users. Slightly related is recent work from the data mining community which leverages social tags from sites like Flickr to create "folksonomies" of concepts [53]. While utilizing data provided by several users, the resulting concept hierarchy is not personalized to any particular one.

Finally, we find that the closest existing work that models concepts over time is the body of work in the topic modeling literature on incorporating time into the models [4, 62]. While these models assume time is a dimension that should be modeled in a corpus, they do not generally admit new documents very easily. Recent work, however, has included streaming topic models, focused on efficient inference techniques that make it easy for a topic model to incorporate new documents [10].

1.2 Modeling and Optimization

Assuming that we are able to represent a document collection with an appropriate set of concepts, how do we encode a particular retrieval task? Specifically, what can we do better now that concepts are defined at an appropriate level of granularity? Answering these questions is important to demonstrating the effectiveness and utility of our proposed concept modeling approaches. While any particular retrieval task will naturally require some level of task-specific modeling, we hope to take advantage of commonly desired properties of result sets to make our methodology generally applicable.

In particular, many retrieval tasks benefit from solutions exhibiting the following characteristics:

- *Result set diversity* is important because redundant documents often add little value.
- *Efficient optimization* is critical because real-world retrieval settings require responsiveness on large document collections.
- *Theoretical guarantees on solution quality* provide a foundation for effective performance on real data.
- *Flexibility in task formulation* can lead to a general approach that solves many different retrieval tasks.

Our initial work has shown that submodular coverage function optimization is a particularly promising way to achieve all the desired characteristics listed above [22, 23]. Specifically, given an appropriate

concept representation, a particular retrieval task can be formulated as the selection of a set of documents that "cover" the concepts of interest. Thus, for a given retrieval task, a *coverage function* must be defined that relates individual documents to the concepts of the collection. For example, in the simplest case, we might say that a document covers a concept if it *contains* that concept. A probabilistic extension to this simple case allow us to assign a probability that a particular document covers a particular concept. We might then select a set of documents that has the highest probability of covering the entire concept space.

In this thesis, we wish to investigate how effective a submodular coverage approach can be for solving a myriad of information retrieval problems.

Related Work Early work by Carbonell et al. [11] introduced the concepts of "relevant novelty" and "marginal relevance" for retrieving diverse result sets in response to a user query. A line of related research is the area of subtopic retrieval [14, 65], where the task is to retrieve documents that cover many subtopics of the given query. In the traditional information retrieval setting, it is assumed that the relevance of each document is independent of the other documents. However, in subtopic retrieval the utility of a document is contingent on the other retrieved documents. In particular, a newly retrieved document is relevant only if it covers subtopics other than the ones covered by previous documents. Thus, the concept of relevance in subtopic retrieval is similar to our notion of coverage, which has a diminishing returns characteristic. However, while subtopic retrieval is query-based, we intend to apply our framework to more general retrieval tasks, even those with no explicit query.

More recent work has investigated the use of structural support vector machines (SVMs) to predict diverse subsets of documents by directly training for subtopic diversity [64]. Such an approach requires labeled training data indicating subtopic coverage, which may be difficult or costly to obtain in practice. More closely related to our approach is the line of recent research on using submodular coverage functions to solve various sensor selection and placement problems, including which blogs a reader should follow to keep up to speed with breaking news [38, 39, 42].

1.3 Interaction

User interaction is an important facet of any information retrieval system. Whether for specifying an information need or providing feedback on selected documents, facilitating and modeling user interaction goes hand in hand with the representational and optimization issues described above.

When the retrieval task involves responding to a user query, as is commonly the case, interaction becomes critical, as the query must be specified in some manner. While a user's complex information need may be best described in a rich, structured manner, this structure is commonly thrown away when trying to specify the need as a keyword query, as is traditionally required. One goal of our research is to ensure that users can specify their needs in the most natural of ways. Thus, freed from the shackles of having to specify every information request as a keyword query, users will be able to discover documents that are more closely related to what they had in mind, and do so more efficiently.

In this thesis, we intend on investigating novel input interactions for users, augmenting or replacing traditional keyword queries in certain settings, potentially with sets, hierarchies or graphs of documents.

In other cases, user interaction is motivated by *personalization*. Ambiguous queries, for instance, are better served when some context is known regarding the user's interests. A frequently used example is the query "jaguar," which could refer to the animal, the luxury car or the electric guitar. Inferring that a user is an avid musician from her interaction with previously retrieved results (e.g., clicks or explicit feedback) would allow a retrieval system to tailor its behavior accordingly.

However, while user feedback can lead to improved results, a key trade-off exists between the amount of user involvement and the amount and quality of information gleaned from the user. Just as we wish to allow the user to specify his or her input most naturally, we seek to obtain the most beneficial user feedback possible in a way that is not onerous to the user. In particular, we intend on accomplishing this by tying personalization to the concept representation and modeling discussed in the previous sections.

In this thesis, we plan on addressing the following two questions:

- **How do we know how much a user is interested in a particular concept, and what happens when that interest changes over time?**
- **How much does a user trust a particular document as it pertains to a given concept?**

With answers to these questions, we can design models (e.g., of the submodular coverage variety discussed in the previous section) that take into account user interests and trust preferences. By estimating user interests for concepts, the fundamental units of information in a document collection (as opposed to associating interests with items such as documents), we hope for a more flexible representation that allows user feedback to better generalize to other sets of documents.

While answers to the first question allow us to resolve ambiguous queries, as described above, we can use answers to the second question to bias document selection towards documents that a particular user trusts. For example, this is an especially important problem in dealing with information overload in scholarly literature, as different scientists trust different authors and publication venues for different scientific concepts.

Furthermore, we note that we certainly do not want to ask a user to give us an explicit trust and interest rating for each of a potentially large set of concepts. Rather, to address both questions, we seek to investigate the notion of *transferred feedback*, where we take data that a user has already generated in one setting (e.g., a user's Twitter feed or BibTeX file), and use it to inform the personalization in another (e.g., selecting news articles or trustworthy research papers).

Finally, given the prominence of personalization in this thesis, and its ubiquity in today's Web experience (e.g., from shopping to search to news feeds), **we wish to investigate whether *overpersonalization* exists, and if so, study its consequences.**

Related Work We are not the first to address the shortcomings of the traditional keyword search paradigm; researchers in both the library science and computer science communities have looked at more natural methods of meeting a user's information need, both among scholarly literature [3, 7, 44, 61] and on the Web [50, 52]. We wish to expand upon this prior work by designing a variety of models that allow for users to move beyond keyword search and specify their information needs in the most natural way possible, while tying into our general coverage framework.

Two common approaches to personalization are *collaborative filtering* [16, 43] and *content-based filtering*. In collaborative filtering, user preferences are learned in a content-agnostic manner by correlating the

user's past activity with data from the entire user community. In a content-based approach, documents are recommended to a user if they are similar to documents that the user previously liked, where similarity is based on document content. Traditional collaborative filtering tends to requires a large user-base for the techniques to be effective, but we hope our notion of transferred feedback that we propose will allow us to leverage *implicit* preference and trust data from many users.

Previous work on determining trust in a document has generally not addressed the issues of personalization. For instance, while a long line of prior work exists in bibliometrics on summarizing the impact of an author or publication with a single number [1], often based on citation statistics [25, 33] or eigenvector methods [15, 36, 51, 54], here we wish to capture a more detailed picture of the trust relationship between a particular user and the authors in a corpus.

Beyond this more general work on finding the authoritativeness of a document, there has been some recent work on trust-based personalization, particularly in the context of recommendation systems (cf. [2]). In contrast to these approaches, we wish to determine trust at a *concept-specific* level, as it is rare for a user to trust an author or document universally, but rather more common for the user to trust someone with respect to a particular idea or concept. Likewise, similar to our notion of transferred feedback is the idea proposed by Teevan et al. for using information from outside the corpus of interest to inform relevance feedback [58]. Moreover, an important paper by Guha et al. describes a framework for how trust and distrust propagates in a social network (and also contains a useful survey of trust as it appears in fields beyond machine learning and computer science) [28]. We intend on building upon this prior work and expanding it to fit our notion of *concept-specific* interests and trust.

To address changing user interests over time, we will have to explore approaches for addressing concept drift in online learning (e.g., [37]). Overpersonalization, on the other hand, has been studied in primarily non-computational settings (e.g., behavioral studies or post-hoc analyses of voting records) (cf. [57]), and we intend on deploying larger-scale studies that attempt to directly and quantitatively measure overpersonalization behavior on the Web.

1.4 Thesis Statement and Research Questions

The core of this thesis research revolves around the following statement:

> By modeling document collections with an appropriate concept representation, we can effectively deal with information overload by framing efficient optimization problems that produce personalized solutions to common retrieval tasks with minimal burden on the user.

In particular, we plan on addressing the following specific research questions over the course of this thesis:

(Q1) What is an appropriate level of concept granularity for a particular retrieval task, document collection and user?

(Q2) How do the concepts in a document collection change as new documents are added?

(Q3) How can a submodular coverage approach be used to solve a variety of retrieval problems?

(Q4) What are appropriate mechanisms with which users can express their information needs?

(Q5) How do we know how much a user is interested in a particular concept, and what happens when that interest changes over time?

(Q6) How much does a user trust a particular document as it pertains to a given concept?

(Q7) Does overpersonalization exist, and if so, what are its consequences, and how can it be addressed?

Some of these questions have been initially addressed with our prior contributions, while others have yet to be investigated. Chapter 2 and Chapter 3 describe our prior and proposed contributions, respectively. Chapter 4 describes a detailed timeline of completion.

Chapter 2

Prior Contributions

2.1 Turning Down the Noise in the Blogosphere (Q3, Q5)

Over the last several years, the blogosphere has experienced a substantial increase in the number of posts published daily[1], and one immediate consequence is that many readers now suffer from information overload. In this joint work with Gaurav Veda and Dafna Shahaf [23], we attempt to *turn down the noise* in the blogosphere by framing this information retrieval problem as probabilistic max-cover over a set of concepts representing the important stories discussed in the blogosphere.

2.1.1 Coverage (Q3)

Figure 2.1(a) shows a typical day in the blogosphere (January 17, 2009). The size of a word is proportional to its frequency across the blogosphere. Examining the picture, we can spot some of the popular stories for that day: the inauguration of Barack Obama and the Israel-Gaza conflict.

Many posts talk about the same story, e.g., the inauguration. Moreover, stories may have a certain degree of overlap. Intuitively, our goal is to select a small set of blog posts that captures the important stories of the day, while at the same time avoiding redundancy.

Concept definition

We can characterize posts in the blogosphere by concepts. In this work, we considered two simple concept representations at opposite ends of the granularity spectrum:

- named entities and noun phrases (fine-grained),
- topics from a latent Dirichlet allocation (LDA) topic model (coarse-grained).

In this initial work, we did not attempt to learn an optimal concept granularity, but rather wished to experiment with the most basic representations. Given a definition of concepts, we can formally define the blogosphere:

[1] From Technorati, `http://technorati.com/state-of-the-blogosphere/`

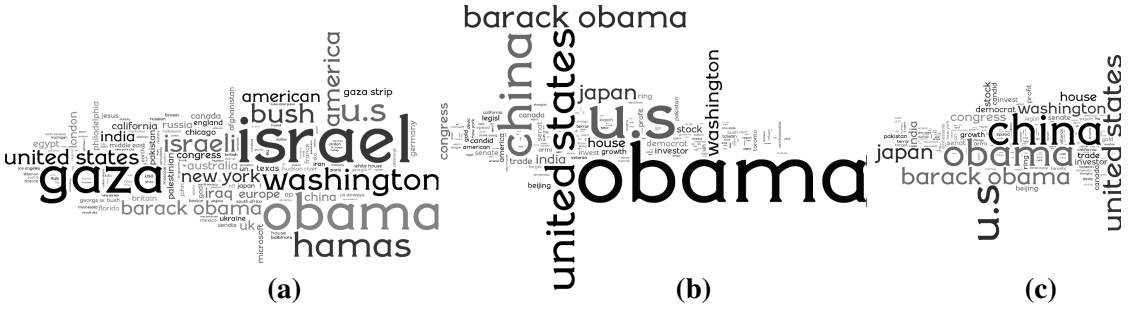


Figure 2.1: (a) Global word frequency across the blogosphere (January 17, 2009). The size of a word is proportional to its frequency. (b) Coverage vs. (c) incremental coverage of a post about Obama and China, given that we already saw a post about Obama. The incremental coverage of Obama is much smaller than the regular coverage.

Definition 2.1.1 (Blogosphere)**.** *A blogosphere is a triplet* $\langle \mathcal{U}, \mathit{Posts}, \mathit{cover}_{\cdot}(\cdot) \rangle$. $\mathcal{U} = \{u_1, u_2, ...\}$ *is a finite set of concepts, and Posts is a finite set of posts. The relation between posts and concepts is captured by the* covering *function.* $\mathit{cover}_j(i) : \mathcal{U} \rightarrow \mathbb{R}^+$ *quantifies the amount post*$_j \in$ *Posts covers concept* u_i.

In the simplest case, $\mathit{cover}_{\cdot}(\cdot)$ is a binary indicator function, turning posts into subsets of concepts. Later, we explore other softer notions of coverage functions, e.g., ones with probabilistic interpretations.

Covering Concepts

Given our model $\langle \mathcal{U}, \mathit{Posts}, \mathit{cover}_{\cdot}(\cdot) \rangle$, we wish to determine how effectively a given small set of posts can cover the important stories in the blogosphere. More formally, our goal is to pick a set of k posts $\mathcal{A} \subseteq \mathit{Posts}$, in order to maximize some coverage objective.

We first consider the well-known *Budgeted Maximum Coverage* problem:
Definition 2.1.2 (Budgeted Maximum Coverage)**.**
Given a set of ground elements $\mathcal{U}$*, a collection* $\mathbb{S} = \{S_1, ...S_m\}$ *of subsets of* $\mathcal{U}$*, and a budget* $k \geq 0$*, select* $\mathcal{A} \subseteq \mathbb{S}$ *of size at most* k *which maximizes the number of covered elements,* $|\bigcup_{S_j \in \mathcal{A}} S_j|$.

In our setting, this coverage can be formalized as maximizing:

$$F(\mathcal{A}) = \sum_{i \in \mathcal{U}} \mathbf{1}(\exists a_j \in \mathcal{A} : \ \mathit{cover}_j(i) = 1),$$

where $\mathbf{1}(\cdot)$ is the indicator function.

Although max-coverage is an NP-hard problem, there are several efficient and effective approximation algorithms for this task. However, this naïve approach suffers from some serious drawbacks:

- *Concept significance in corpus*: All concepts in a corpus are treated equally, and thus we cannot emphasize the importance of certain concepts. For example, covering "Cathedral High School" should not be as valuable as covering "Obama."
- *Concept significance in post*: This objective function does not characterize how relevant a post is to a particular concept, e.g., a post about Obama's speech covers Obama just as much as a post that barely mentions him. As a side effect, this objective rewards "name-dropping" posts (posts that include many concepts, without being about any of them).

- *Incremental coverage*: This coverage notion is too strong, since after seeing one post that covers a certain concept, we will never gain anything from another post that covers the same concept. This does not correspond to our intuitive notion of coverage, which should be subject to the law of diminishing returns: each additional time we see a concept we get an additional reward, which decreases with the number of occurrences. For example, suppose we show the user a post about Obama's inauguration. The second post we consider showing her is about the effect of Obama's presidency on China. Figure 2.1(b) shows the raw coverage of the second post, and "Obama" is the top-covered concept. However, if we take into account the fact that we have already covered the concept "Obama" to some extent by the first post, the coverage by the second post changes. Figure 2.1(c) shows the *incremental coverage* by the second post. As illustrated, the significance of this post towards "Obama" is diminished, and most of our reward would come from covering "China."

We now address each of these three issues.

To address *Concept significance in corpus*, we can simply assign weights w_i to each concept u_i:

$$F(\mathcal{A}) = \sum_{i \in \mathcal{U}} w_i \, \mathbf{1}(\exists a_j \in \mathcal{A} : \; \textit{cover}_j(i) = 1).$$

If concepts are words, the weights might correspond to their frequency in the data set.

We now turn our attention to *Concept significance in post.* Each post should exhibit different degrees of coverage for the concepts it contains, which can be achieved by softening the notion of coverage, $\textit{cover}_j(i)$. One approach is to use a generative model to estimate the probability of a concept given a post, $P(u_i \mid \textit{post}_j)$. If, for example, our concepts are topics discovered by a topic model, then this term is simply the probability that document j is about topic i. More generally, any generative model for the particular set of concepts can be used to define this probability.

Given such a probabilistic model, we can define the notion of soft coverage more formally. If our concepts are sufficiently coarse-grained, e.g., topics in a topic model, then a post can be thought of as being about a single concept, in which case $\textit{cover}_j(i) = P(u_i \mid \textit{post}_j)$. Alternatively, for lower-level concepts, such as named entities, we could assume that each post is about ℓ concepts. If these concepts are picked, for example, at random with replacement from $P(u_i \mid \textit{post}_j)$, then our coverage will become $\textit{cover}_j(i) = 1 - (1 - P(u_i \mid \textit{post}_j))^{\ell}$. By requiring that all posts cover the same number of concepts, we alleviate the problem of "name-dropping," since a post cannot cover a large number of concepts well.

The probabilistic approach allows us to define concept importance in individual posts as well as in the whole corpus. However, if we define coverage as $F(\mathcal{A}) = \sum_{i \in \mathcal{U}} w_i \; \sum_{a_j \in \mathcal{A}} \textit{cover}_j(i)$, then the *Incremental coverage* problem would persist, as this function does not possess the diminishing returns property. Instead, extending the probabilistic interpretation further, we can view set-coverage as a sampling procedure: each post tries to cover concept i with probability $\textit{cover}_j(i)$, and the concept is covered if at least one of the posts in $\mathcal{A}$ succeeded. Thus, as $\mathcal{A}$ grows, adding a post provides less and less additional coverage. Formally, we can define the probabilistic coverage of a concept by a set of posts $\mathcal{A}$ as:

$$\textit{cover}_{\mathcal{A}}(i) = 1 - \prod_{a_j \in \mathcal{A}} (1 - \textit{cover}_j(i)). \tag{2.1}$$

Finally, we propose the following objective function for the problem of probabilistic coverage of the blogosphere:

$$F(\mathcal{A}) = \sum_{i \in \mathcal{U}} w_i cover_{\mathcal{A}}(i). \tag{2.2}$$

Our task is to find k posts maximizing the above objective function:

$$\mathcal{A}^* = \underset{\mathcal{A} \subseteq Posts: |\mathcal{A}| \leq k}{\operatorname{argmax}} F(\mathcal{A}). \tag{2.3}$$

Optimizing Coverage of the Blogosphere

Using the notion of coverage in Eq. 2.1, our goal now is to find the set of posts $\mathcal{A}$ that maximizes our objective function in Eq. 2.2. Unfortunately, we can show by reduction from max-coverage that this objective is NP-complete, suggesting that the exact maximization of this function is intractable. However, our objective function satisfies an intuitive diminishing returns property, *submodularity*, which allows us to find good approximations very efficiently:

Definition 2.1.3 (Submodularity)**.** *A set function F is* submodular *if,* $\forall \mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}, \forall s \in \mathcal{V} \setminus \mathcal{B}, F(\mathcal{A} \cup \{s\}) - F(\mathcal{A}) \geq F(\mathcal{B} \cup \{s\}) - F(\mathcal{B})$.

Claim 2.1.4. *The probabilistic coverage function for the blogosphere in Eq. 2.2 is submodular [23].*

Intuitively, submodularity characterizes the notion that reading a post s after reading a small set of posts $\mathcal{A}$ provides more coverage than reading s after having already read the larger set $\mathcal{B} \supseteq \mathcal{A}$.

Although maximizing submodular functions is NP-hard [35], by discovering this property in our problem, we can take advantage of several efficient approximation algorithms with theoretical guarantees. For example, the classic result of Nemhauser et al. [49] shows that by simply applying a greedy algorithm to maximize our objective function in Eq. 2.2, we can obtain a $(1 - \frac{1}{e})$ approximation of the optimal value. Thus, a simple greedy optimization can provide us with a near-optimal solution. However, since our set of posts is very large, a naïve greedy approach can be too costly. Therefore, we use CELF [42], which provides the same approximation guarantees, but uses lazy evaluations, often leading to dramatic speedups.

2.1.2 Personalization (Q5)

Thus far, we have defined a global notion of coverage for the blogosphere. However, each user has different interests, and the selected posts that cover the prevalent stories may contain many topics that do not interest him. Instead, our goal in this section is to utilize user feedback in order to *learn* a personalized notion of coverage for each user.

Recall that, in the previous section, $F(\mathcal{A})$ assigns a fixed weight w_i to every concept, representing its importance. In practice, concept importance varies among different users. One user might care about a concept "NASCAR," while others may be indifferent to it. To address this issue, we augment the fixed

weights w_i with personalized preferences π_i for each concept i. In the following, we assume that a user's coverage function is of the form:

$$F_{\pi^*}(\mathcal{A}) = \sum_{i \in \mathcal{U}} \pi_i^* \; w_i \; \textit{cover}_{\mathcal{A}}(i), \tag{2.4}$$

for some unknown set of weights $\{\pi_i^*\}$. Our goal now is to learn a user's coverage function $F_{\pi^*}(\mathcal{A})$ by learning this optimal set of preferences $\{\pi_i^*\}$.

Interaction Models

In order to receive personalized results, users need to communicate their preferences. Since F_π is a set function, the most natural notion of feedback from a machine learning perspective would be for users to provide a single label for the set of posts that they are presented, indicating whether they like or dislike the entire set. However, this approach suffers from two limitations. First, from the point of view of the user, it is not very natural to provide feedback on an entire set of posts. Second, since there are exponentially many such sets, we are likely to need an extensive amount of user feedback (in terms of sets of posts) before we could learn this function. Instead, we assume that users go through a list of posts $\mathcal{A}$ in order, submitting feedback f_j ("liked"= +1, "indifferent" = 0, "disliked" = -1) for each post $a_j \in \mathcal{A}$. We take no feedback on a post to mean "indifferent."

Personalization by Minimizing Regret

Our objective function is defined in terms of sets, but our feedback is in terms of individual posts. How should we provide an appropriate credit assignment?

To address this issue, we consider the *incremental coverage* of a post, i.e., the advantage it provides over the previous posts. The incremental coverage we receive by adding post a_j to the set $\mathcal{A}$ is:

$$\textit{inc-cover}_j(\mathcal{A}, i) = \textit{cover}_{\mathcal{A} \cup a_j}(i) - \textit{cover}_{\mathcal{A}}(i).$$

Note that if $\textit{cover}_{\mathcal{A}}(i)$ is defined as in Eq. 2.1, then the incremental coverage is the probability that a_j is the first post to cover concept u_i. Furthermore, if we view the set of documents $\mathcal{A}$ as an ordered set $\mathcal{A} = \{a_1, \ldots, a_k\}$[2], the sum of incremental coverages is a telescoping sum that yields the coverage of a set of documents $\mathcal{A}$:

$$\begin{aligned} \sum_{a_j \in \mathcal{A}} \textit{inc-cover}_j(a_{1:j-1}, i) &= \sum_{a_j \in \mathcal{A}} \textit{cover}_{a_{1:j}}(i) - \textit{cover}_{a_{1:j-1}}(i), \\ &= \textit{cover}_{\mathcal{A}}(i), \end{aligned}$$

where $a_{1:j-1}$ is shorthand for the set of documents $\{a_1, \ldots, a_{j-1}\}$.

Using incremental coverages, we can now define the reward we receive after presenting $\mathcal{A}$ to a user with preferences π and obtaining feedback f:

$$\textit{Rew}(\pi, \mathcal{A}, f) = \sum_{i \in \mathcal{U}} \pi_i \; w_i \sum_{a_j \in \mathcal{A}} f_j \;\; \textit{inc-cover}_j(a_{1:j-1}, i).$$

[2]This ordering could be defined by the order the posts are presented to the user, e.g., the one picked by the greedy algorithm.

If the user liked all of the documents in $\mathcal{A}$ (i.e., $\forall j, f_j = 1$), this reward becomes exactly the coverage function we are seeking to maximize, $F_\pi(\mathcal{A}) = \sum_{i \in \mathcal{U}} \pi_i \, w_i \, cover_{\mathcal{A}}(i)$, as in Eq. 2.4.

Our algorithm maintains an estimate of the user's preferences at each time step t, $\pi^{(t)}$. Given this estimate, we optimize $F_{\pi^{(t)}}(\mathcal{A})$ and pick a set of documents $\mathcal{A}^{(t)}$ to show the user. After receiving feedback $f^{(t)}$, we gain a reward of $Rew(\pi^{(t)}, \mathcal{A}^{(t)}, f^{(t)})$. After T time steps, our average reward is therefore:

$$AvgRew(T) = \frac{1}{T} \sum_{t=1}^{T} Rew(\pi^{(t)}, \mathcal{A}^{(t)}, f^{(t)}).$$

Since our decisions at time t can only take into account the feedback we have received up to time $t-1$, the decisions we made may have been suboptimal. For comparison, consider the reward we would have received if we had made an informed choice for the user's preferences π considering all of the feedback from the T time steps:

$$BestAvgRew(T) = \max_{\pi} \frac{1}{T} \sum_{t=1}^{T} Rew(\pi, \mathcal{A}^{(t)}, f^{(t)}). \tag{2.5}$$

That is, after seeing all the user feedback, what would have been the right choice for user preference weights π? The difference between our reward and this best choice in retrospect is called the *regret*:

Definition 2.1.5 (Regret)**.** *Our average regret after T time steps is the difference $BestAvgRew(T) - AvgRew(T)$.*

Positive regret means that we would have preferred to use the weights π that maximize Eq. 2.5 instead of our actual choice of weights $\pi^{(t)}$. A *no-regret learning algorithm*, such as the one we describe in the next section, will allow us to learn $\pi^{(t)}$ such that, as T goes to infinity, the regret will go to zero at a rapid rate. Intuitively, this no-regret guarantee means that we learn a sequence $\pi^{(t)}$ that does as well as any fixed π–including the true user preferences, π^*–on the sets of posts that the user is presented. By learning the personalized coverage function for a particular user in this manner, the posts we provide will be tailored to his tastes.

Learning a User's Preferences

We now describe our algorithm for learning π^* from repeated user feedback sessions. Like many online algorithms [12], our approach updates our estimated $\pi^{(t)}$ using a multiplicative update rule. In particular, our approach can be viewed as a special case of Freund and Schapire's multiplicative weights algorithm [24].

The algorithm starts by choosing an initial set of weights $\pi^{(1)}$. (WLOG, we assume weights are normalized to sum to 1, since the coverage function is insensitive to scaling.) In the absence of prior knowledge about the user, we can choose the uniform distribution:

$$\pi_i^{(1)} = \frac{1}{|\mathcal{U}|}.$$

If we have prior knowledge about the user, we can start from the corresponding set of weights.

At every round t, we use our current distribution $\pi^{(t)}$ to pick k posts, $\mathcal{A}^{(t)}$, to show the user. After receiving feedback $f^{(t)}$, we would like to increase the weight of concepts covered by posts the user liked,

and decrease the weight of concepts covered by posts the user disliked. These updates can be achieved by a simple multiplicative update rule:

$$\pi_i^{(t+1)} = \frac{1}{Z}\, \pi_i^{(t)}\, \beta^{-\mathcal{M}(i,f^{(t)})}, \tag{2.6}$$

where Z is the normalization constant, $\beta \in (0,1)$ is the learning rate, and, intuitively, $\mathcal{M}(i, f^{(t)})$ measures the contribution (positive or negative) that feature i had on our reward:

$$\mathcal{M}(i, f^{(t)}) := \frac{w_i \sum_{a_j \in \mathcal{A}^{(t)}} f_j^{(t)}\ \textit{inc-cover}_j(a_{1:j-1}, i)}{2 \max_i w_i}, \tag{2.7}$$

where the normalization by $2 \max_i w_i$ is simply used to keep this term in the range $[-0.5, 0.5]$.

If the learning rate β is small, we make large moves based on the user feedback. As the learning rate tends to 1, these updates become less significant. Thus, intuitively, we will start with a small value of β and slowly increase it.

Claim 2.1.6. *If, for number of personalization epochs T, we use a learning rate β_T given by:*

$$\beta_T := \frac{1}{1 + \sqrt{\frac{2 \ln |\mathcal{U}|}{T}}}, \tag{2.8}$$

then our preference learning procedure will have regret bounded by:

$$BestAvgRew(T) - AvgRew(T) \leq \mathcal{O}\left(\sqrt{\frac{\ln |\mathcal{U}|}{T}}\right).$$

Since our regret goes to zero as T goes to infinity, our approach is called a no-regret algorithm. The proof follows from Freund and Schapire [24], by formalizing our learning process as a two-player repeated matrix game involving our algorithm and the user.

2.1.3 Evaluation

We evaluate our algorithm on real blog data collected over a two week period in January 2009. These posts come from a diverse set of blogs, including personal blogs, blogs from mainstream news sites, commercial blogs, and many others.

We obtain the data from Spinn3r, collecting approximately 500,000 posts per day. After performing some simple data cleaning steps, such as removing web forums and classifieds, we reduce this number to about 200,000 posts per day in our data set. However, as this is real Web data, it is still invariably noisy even after cleaning. Thus, our algorithm must be robust to content extraction problems.

We implemented two variants of our algorithm, Turn Down the Noise (TDN), and we refer to the variant of our algorithm with coarse-grained concepts as TDN+LDA and the variant with fine-grained concepts as TDN+NE.

Evaluating Coverage

As detailed previously, the main objective of our algorithm is to select a set of posts that best covers the important and prevalent stories currently being discussed in the blogosphere. The major world events that took place during the time corresponding to our data set included the Israel-Gaza conflict, the inauguration of Barack Obama, the gas dispute between Russia and Ukraine, as well as the global financial crisis. As an example, here is the set of posts that our algorithm selects for an eight hour period on January 18, if our budget k is set to five:

1. Israel unilaterally halts fire as rockets persist
2. Downed jet lifted from ice-laden Hudson River
3. Israeli-trained Gaza doctor loses three daughters and niece to IDF tank shell
4. EU wary as Russia and Ukraine reach gas deal
5. Obama's first day as president: prayers, war council, economists, White House reception

The selected five posts all cover important stories from this particular day. The Israel-Gaza conflict appears twice in this set, due to its extensive presence in the blogosphere at the time. It is important to note, however, that these two posts present different aspects of the conflict, each being a prevalent story in its own right. By expanding the budget to fifteen posts, the algorithm makes additional selections related to other major stories of the day (e.g., George W. Bush's legacy), but also selects "lifestyle" posts on religion and cooking, since these represent the large portion of the blogosphere that is not directly related to news and current events.

As another example, here are the top five selected posts from the morning of January 23, the day after the Academy Award nominations were announced:

1. Button is top Oscar nominee
2. Israel rules out opening Gaza border if Hamas gains
3. Paterson chooses Gillibrand for U.S. Senate
4. Fearless Kitchen: Recipe: Medieval Lamb Wrap
5. How Obama avoided a misguided policy blunder

A post describing the Oscar-nominated movie *The Curious Case of Benjamin Button* supplants the Israel-Gaza conflict at the top of the list, while a cooking post makes it up to the fourth position.

We wish to quantitatively evaluate how well a particular post selection technique achieves the notion of coverage we describe above on *real blog data*. However, the standard information retrieval metrics of precision and recall are not directly applicable in our case, since we do not have labels identifying all the prevalent stories in the blogosphere on a given day and assigning them to specific posts. Rather, we measure the *topicality* of individual posts as well as the *redundancy* of a set of posts. We say a post is *topical* with respect to a given time period if its content is related to a major news event from that period. A post r is *redundant* with respect to a previous post p if it contains little or no additional information to post p. An ideal set of posts that covers the major stories discussed in the blogosphere would have high topicality and low redundancy.

We conducted a study on 27 users to obtain labels for topicality and redundancy on our data. We compared TDN+LDA and TDN+NE to four popular blog aggregation sites: the front page of Digg, Google Blog Search, Nielsen BuzzMetrics' BlogPulse, and Yahoo! Buzz.

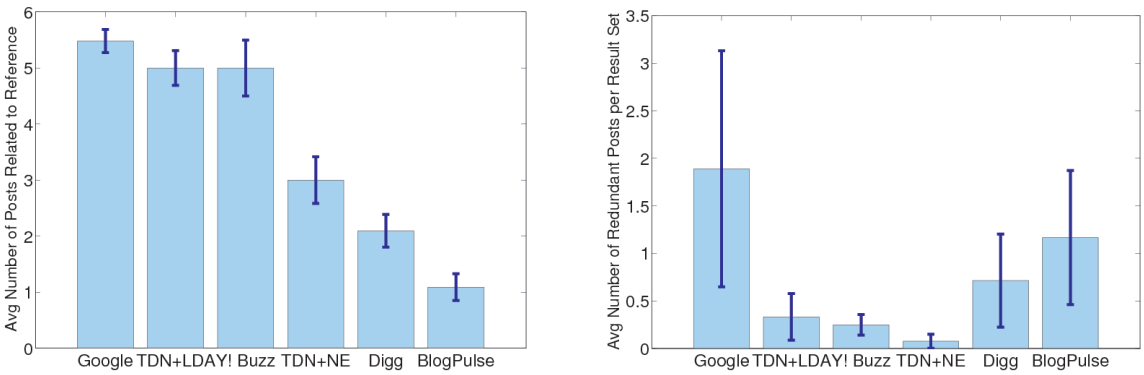


Figure 2.2: Left: Results from user study measuring topicality. The bars show the average number of posts (out of 10) that users found to be topical with respect to the reference stories. Right: Results of the redundancy user study. Users report the number of redundant posts for each post selection technique they are presented with. Error bars on all plots indicate standard error.

Measuring Topicality In order for users to measure the topicality of a blog post, they need an idea of what the major news stories are from the same time period. We express this information to our study participants by providing them with headlines gathered from major news sources in five different categories: world news, politics, business, sports, and entertainment. The headlines for each category are aggregated from three different news sources to provide a wider selection for the users and to avoid naming a single source as the definitive news outlet for a category. For instance, for politics we present headlines from Reuters, *USA Today*, and *The Washington Post*. This collection of headlines is akin to a condensed newspaper, and we refer to these stories as *reference stories*.

We present the participants with reference stories gathered at a particular time, e.g., January 18, 2009, 2:00pm EST, which we call the *reference time*. We then show each participant a set of ten posts that was chosen by one of the six post selection techniques, and ask them to mark whether each post is "related" to the reference stories. Each post is presented as a title along with a short description. The users are not made aware of which technique the posts come from, so as not to bias their ratings. The posts selected by TDN+LDA and TDN+NE were chosen from an eight hour window of data ending at the reference time, while the posts selected by the popular blog aggregation sites were retrieved from these sites within fifteen minutes of the reference time.

Figure 2.2(left) shows the results of the topicality user ratings on the six techniques. On average, the sets of ten posts selected by Google Blog Search, TDN+LDA and Yahoo! Buzz each contain five topical posts out of ten presented. The topicality of these techniques is significantly better than that of TDN+NE, Digg and BlogPulse. BlogPulse selects the most linked-to posts of the day, which does not seem to be a good heuristic for covering the important stories. Many of these posts are technology how-to pages, such as "Help With Social Bookmarking Sites," the highest ranked post from January 18. Digg selects its top posts by user voting, and thus the top selected posts consist of a few prevalent stories and many entertaining or shocking posts, such as "Teen Stabbed But Makes It To Job Interview," the top post from February 6.

TDN+LDA outperforms TDN+NE because high-level concepts, such as LDA topics, capture stories in a better way than low-level features do. For example, for one eight hour period in our data set, there is a

coherent LDA topic about the EU-Russia gas crisis. Therefore, when we cover this topic, we will present a story that is about the crisis. However, the named entity "Russia" may be covered by multiple stories. TDN+NE selects a post about Russia's plan to go ahead with the opening of a pediatric medical center in Moscow despite the current financial crisis, since it contains important named entities and nouns like "Russia," "Putin," "crisis," etc. Hence, if we only cover low-level features, we might select a post that is not topical, yet contains multiple important features.

While topicality captures a major aspect of our notion of coverage, in that important current events are covered by the selected posts, one drawback of this evaluation method is that lifestyle blog posts are not adequately represented. It is difficult to define a set of reference sites that summarize the day's most important recipes or most prevalent do-it-yourself tips, for instance. Furthermore, in our case, we did not want to show our study participants more than five categories of reference stories, so as not to overwhelm them. As a result, a post related to an important technology story would likely not be considered topical, as we left this category out.

Measuring Redundancy The user study described in the previous section allowed us to measure whether posts were topical or not. However, topicality is not enough to judge the goodness of a set of posts, since they may all be about the same story, and hence not interesting. Instead, we want the posts to be diverse, so that they capture all of the important stories in the blogosphere, as well as appeal to everyone's interests. As part of our user study, we asked users to look at a set of fifteen posts selected by one of the six previously described post selection techniques, and mark any occurrences they thought were redundant. Each of 27 participants was presented with either two or three sets of posts generated by different algorithms over the same time period. The users were not aware of the sources of the posts.

Figure 2.2(right) shows that both variants of our algorithm outperform Digg, BlogPulse and Google Blog Search on the redundancy metric. In other words, our algorithm selects diverse sets of posts. This diversity is primarily due to the diminishing returns property of our objective function. If we have covered the important features of a story once, covering it again yields only a small reward. Google Blog Search has the highest number of redundant results, and has high variance, suggesting that on some days many of the posts on its front page are similar. In fact, on average, the posts selected by Google Blog Search are nearly six times as redundant as those selected by TDN+LDA.

However, it should be noted that performing well on the redundancy metric alone is not sufficient. For example, it may turn out that all the posts picked by an algorithm are non-redundant, but meaningless, and hence of no interest to a user. Thus, an algorithm needs to perform well on both the topicality and the redundancy metric in order for it to be useful.

TDN+LDA and Yahoo! Buzz were the two techniques that performed well in both metrics. However, while Yahoo! Buzz uses Web search trends, user voting and other features to select its posts, TDN+LDA achieves the same topicality and redundancy performance by selecting posts only using simple text features.

Personalization

There are two methods by which we evaluate how well our algorithm personalizes the posts it selects in response to user feedback. In one setting, we conduct a user study to directly measure how many of the presented posts a study participant would like to read. In the second setting, we simulate user preferences

on a targeted set of blog posts and observe how our objective function $F(\mathcal{A})$ changes with respect to the unpersonalized case.

Preferences of Real Users We divide our blog data into 33 eight hour segments (epochs), and pick a starting segment at random for a particular user. We present our user with a set of ten posts from his starting segment, selected using TDN+LDA. The posts are displayed as a title and short summary. The user is instructed to read down the list of posts and, one by one, mark each post as "would like to read," "would not like to read," or "indifferent." The user is told to make each decision with respect to the previous posts displayed in that set, so as to capture the notion of incremental coverage. For example, a user might be excited to read a post about Obama's inauguration appearing at the top slot in a particular result set, and thus would mark it as "like to read." However, if four other very similar posts appear below it, by the time he gets to rating the fifth inauguration post in a row, he will likely label it as "not like to read."

After each set of ten posts, our personalization algorithm uses the user ratings to update the weights $\pi^{(t)}$, and selects a personalized set of posts for the next epoch[3]. We also ask the user to mark his preferences on unpersonalized posts presented for the same epochs. The order in which these two conditions are presented is randomized. We repeat this process for a total of five epochs. As this is not a longitudinal study, and we do not wish it to be overly tedious for our participants, we accelerate the personalization process by using a learning rate β of 0.5, corresponding to a short-term learning horizon (i.e., $T \approx 9$ from Eq. 2.8).

Figure 2.3(a) shows the result of this study on twenty users. The vertical axis of the plot shows the average number of posts liked by a user in a single epoch. As one would expect, at epoch 0, when the posts are always unpersonalized, the number of liked posts is approximately the same between the personalized and unpersonalized runs. However, in just two epochs, the users already show a preference towards the personalized results.

If a user only prefers sports posts, personalization is easy, as the user's interests are narrow. In our study, however, the participants were simply instructed to rate posts with their own personal preferences. As people are often eclectic and have varied interests, this task is harder, but more realistic. Thus, it is notable that we are still able to successfully adjust to user tastes in very few epochs, showing a significant improvement over the unpersonalized case.

If instead of asking users to rate posts according to their personal tastes, we ask them to pretend that they only want to read posts on a specific subject (e.g., India), we observe interesting qualitative behavior. Initially, the top posts selected are about the main stories of the day, including the Israel-Gaza conflict and the Obama inauguration. After a few epochs of marking any India-related posts as "like" and all others as "dislike," the makeup of the selected posts changes to include more posts about the Indian subcontinent (e.g., "Pakistan flaunts its all-weather ties with China"). This is particularly notable given that these posts appear relatively infrequently in our data set, and thus without personalization, are rarely selected. Also, while after enough epochs, stories about India eventually supplant the other major news stories at the top of the result set, the Israel-Gaza stories do not disappear from the list, due to their high prevalence. We believe this is precisely the behavior one would want from such a personalization setting.

[3] As topics tend to change from one epoch to the next, we employ a simple bipartite matching algorithm to map personalization weights across epochs. Alternatively, one could use more recent topic models that are designed to work on streaming data [10].

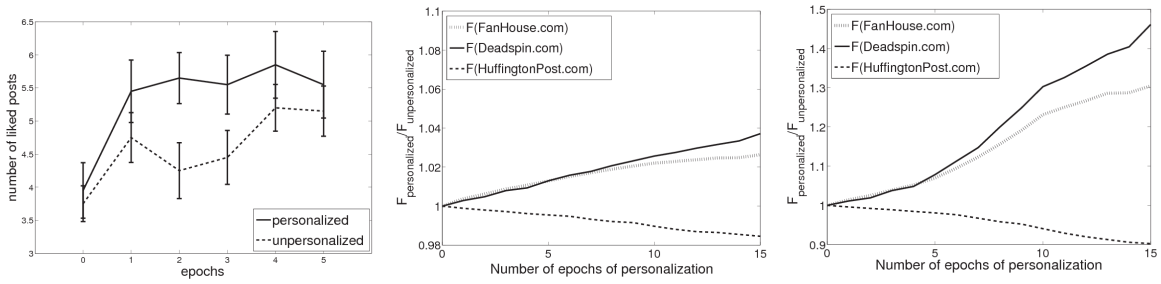


Figure 2.3: (a) Results of the personalization user study, measuring how many posts each user liked, out of the ten presented by TDN+LDA in each epoch. The personalized line corresponds to a learning rate $\beta = 0.5$. (b,c) Effect of number epochs spent personalizing to the simulated preferences of a sports fan on the objective function F, with respect to no personalization. F is evaluated on two sports blogs and one politics blog. Learning rate $\beta = 0.5$ (b), 0.1 (c)

Simulating Preferences We consider the case of a hypothetical sports fan, who always loves to read any sports-related post. In particular, every day, he is presented with a set of posts from the popular sports blog FanHouse.com, and he marks that he likes all of them. We simulate such a user in order to empirically examine the effect of personalization on the objective function.

Specifically, we simulate this sports fan by marking all FanHouse.com posts as "liked" over a specified number of personalization epochs, updating the personalization weights $\pi^{(t)}$at each epoch. On the next epoch, which we call the evaluation epoch, we compute our objective function $F(\mathcal{A})$ on three different sets of posts. First, we compute $F(\mathcal{A})$ on the FanHouse.com posts from this epoch, hypothesizing that the more epochs we spend personalizing prior to the evaluation epoch, the higher this value will be. Second, we compute $F(\mathcal{A})$ on all the posts from DeadSpin.com, another popular sports blog. We also expect to see a higher value of our objective in this case. Finally, we compute $F(\mathcal{A})$ on all the posts from the HuffingtonPost.com Blog, a popular politics blog. The expectation is that by personalizing on sports posts for several days, $F(\mathcal{A})$ for a set $\mathcal{A}$ of politics posts will decrease with respect to the unpersonalized case.

Figure 2.3(b) shows the results of this experiment with a β value of 0.5, and we observe precisely the hypothesized behavior. The vertical axis of this plot shows the ratio of $F(\mathcal{A})$ computed with the learned personalization weights to that of $F(\mathcal{A})$ with the unpersonalized uniform weights, allowing us to compare across the three blogs. Thus, points on the plot that appear higher along the vertical axis than 1 indicate an improvement over the unpersonalized case, while any value below 1 indicates a decline with respect to the unpersonalized case.

Figure 2.3(c) shows the same simulation but with $\beta = 0.1$. This is an aggressive setting of the learning rate, and thus, as expected, the plot shows the objective function changing in the same direction but more rapidly when compared to Figure 2.3(b). These figures capture an important trade off for a deployed system, in that by varying the learning rate β, we trade off the speed of personalization with the variety of selected posts.

2.1.4 Summary

Even with simple concepts, we demonstrated that a probabilistic max-cover approach does an excellent job in retrieving the important stories of the day, especially when compared with the most popular alternatives (Q3). Moreover, we demonstrated an effective, theoretically-sound method for using limited user feedback to determine user preferences for concepts, based on an online learning framework (Q5).

We expect performance on both facets to improve with more nuanced concept selection.

2.2 Beyond Keyword Search: Discovering Relevant Scientific Literature (Q3, Q4, Q6)

A different–but no less daunting–information overload problem occurs in the realm of scholarly research, where scientists conducting literature searches find themselves overwhelmed with millions of articles. In this work, we show that our probabilistic max-cover approach naturally generalizes to more complicated retrieval tasks, and is able to utilize a richer coverage model connecting concepts with documents. While the interested reader can find more details in the full paper [22], here we highlight three aspects of this work that complement what we described in the previous section:

1. Expressing information needs;
2. Complex coverage modeling;
3. Trust modeling.

2.2.1 Expressing information needs (Q4)

Currently, researchers primarily rely on keyword search of online indices such as Google Scholar and PubMed to help them combat information overload. While these tools are indispensable, there are many instances where a researcher's information need cannot be easily specified as a simple string of keywords. Often, such a keyword query is either overly broad, returning many articles that are at best loosely related to the researcher's specific need, or too narrow, potentially returning no articles at all. In these occasions, it may be more natural for the scientist to specify his query as a small set of papers rather than as a set of words. In particular, having already read some articles that are related to the specific task at hand, the scientist can ask, "given that these papers represent my immediate research focus, what else should I read?".

Here, we present an algorithm for discovering relevant scientific literature by responding to queries of this form. More formally, given a small set of papers $\mathcal{Q}$ that we refer to as the *query set*, we seek to return a set $\mathcal{A}$ of additional papers that are related to the concept defined by the query. A paper that cites all of the articles in $\mathcal{Q}$ is likely to represent related research. Likewise, a paper that is cited by every article in $\mathcal{Q}$ might contain relevant background information. However, it is restrictive to require the papers in $\mathcal{A}$ to have a direct citation to or from every article in the query set, as such papers are not guaranteed to exist. Instead, we select a set $\mathcal{A}$ that maximizes a more general notion of *influence* to and from the papers in $\mathcal{Q}$.

2.2.2 Complex coverage modeling (Q3)

In Turning Down the Noise (TDN), we only considered the textual content of the documents when modeling the retrieval problem. However, in scientific literature, the link structure of research papers is too informative to ignore, and thus we must design a more sophisticated coverage model to incorporate both text and citations. We find that we are able to seamlessly extend our max-cover framework to take such rich features into account.

Specifically, for each concept c in our collection of concepts $\mathcal{C}$, we let the probability of influence between two papers x and y with respect to concept c be denoted as $\mathit{Influence}_c(x \leftrightarrow y)$. (Details on how we define this influence can be found in the paper [22].) As we wish to select a *set* of articles that exhibit a high degree of influence to and from the *set* $\mathcal{Q}$ of query documents, we must consider how to combine the influences to and from all the papers in $\mathcal{A}$. By framing this retrieval task in the probabilistic max-cover framework, we can take advantage of the desirable properties of diversity, efficiency, and theoretical guarantees on solution quality that we described previously. As such, we define the influence between a query document q and a set $\mathcal{A}$ as the probability that influence exists between q and *at least one* document in $\mathcal{A}$:

$$\mathit{Influence}_c(q \leftrightarrow \mathcal{A}) \quad = \quad 1 - \prod_{d \in \mathcal{A}} \left(1 - \mathit{Influence}_c(q \leftrightarrow d)\beta_d^{(c)}\right). \tag{2.9}$$

Here, $\beta_d^{(c)}$ is a scalar that can be interpreted as the probability concept c is important in paper d, which safeguards against selecting documents only tangentially related to the important concepts in the query papers.

Given this definition of set influence, we can now define an objective function that, when maximized, returns a diverse set of papers highly relevant to the query:

$$F_{\mathcal{Q}}(\mathcal{A}) \quad = \quad \sum_{q \in \mathcal{Q}} \sum_{c \in \mathcal{C}} \gamma_q^{(c)} \mathit{Influence}_c(q \leftrightarrow \mathcal{A}). \tag{2.10}$$

$\gamma_q^{(c)}$ is a weight proportional to the prevalence of concept c in query document q, ensuring that the influence is focused on the concepts important to the query articles.

We note that the influence functions here plays the same role as the coverage functions in TDN.

2.2.3 Trust modeling (Q6)

While optimizing Equation 2.10 may return articles that are relevant to a particular query, a major shortcoming is that every researcher who submits this query will receive an identical result set. For any given topic, different researchers trust different authors and publications, and the objective in Equation 2.10 provides no means to express these preferences. Moreover, the style of personalization described for TDN does not directly apply here, as a user's trust preferences are not the same as a user's interests.

In order to properly model such an individual notion of *trust* in the setting of scholarly research, we consider two motivating scenarios:

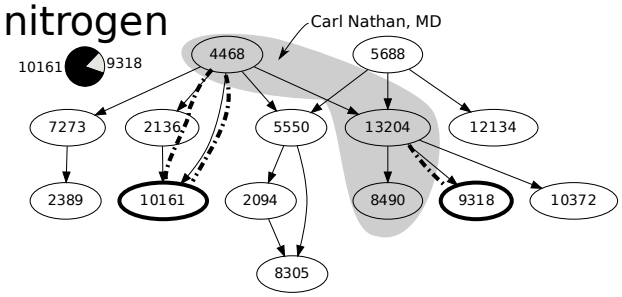


Figure 2.4: Example illustrating trust calculation for an immunologist asking, "How much do I trust Carl Nathan with respect to the concept 'nitrogen'?" Thick dashed lines indicate influence from Dr. Nathan to individual elements of $\mathcal{B}$, and pie chart represents relative prevalence of the word "nitrogen" in the two papers in $\mathcal{B}$.

1. Different authors command different levels of respect from their research communities, e.g., a Nobel laureate versus a first-year graduate student, as an extreme case.
2. Even among distinguished scientists, a particular researcher's interests may be aligned more closely with some than others. Thus, beyond simply differentiating novices from experts, a notion of trust should also capture differences in research interests. For example, asking computer scientists to name whom they most associate with the concept "network" may yield Judea Pearl (Bayesian networks), Jon Kleinberg (social networks), Geoff Hinton (neural networks) or Van Jacobson (computer networks), depending on who is answering. All are distinguished researchers, but each is associated with a distinct subfield of computer science.

At the heart of both scenarios is a personal question that is often answered differently by different researchers: *How much do I trust this author with respect to this concept?*

By answering this question, a researcher would enable us to formally incorporate his trust preferences into our objective function, allowing us to select papers tailored specifically to his tastes. However, as researchers will not be able to provide an answer for every combination of authors and concepts, we must elicit their trust preferences in a less onerous manner. In order to do so, we assume that trust is *transitive*. For example, if Alice trusts an article, and that article is heavily influenced by Bob with respect to the concept "network," then Alice is likely to also trust Bob with respect to "network." Thus, at a fundamental level, a researcher need only specify a set of trusted papers $\mathcal{B}$, from which we can infer answers to the above question. As a shortcut, a researcher may choose to define $\mathcal{B}$ indirectly by specifying a list of trusted journals and conferences, or subsets thereof (e.g., a particular conference track or article classification). $\mathcal{B}$ could also be specified as the papers *cited by* one or more trusted authors, representing a look at one's research through the eyes of another scientist, potentially in another field. Thus, a plucky physicist could ask, "What would Steven Chu recommend I read?", and obtain a set of papers related to his query, yet tailored to the research interests and trust preferences of the Nobel laureate.

With this intuition in mind, we define $\tau^{(c)}_{a|\mathcal{B}}$, the probability a researcher trusts author a with respect to concept c, given trusted articles $\mathcal{B}$. (The "$|\mathcal{B}$" notation in this section indicates personalizing with respect to trusted set $\mathcal{B}$.) Figure 2.4 illustrates how we compute $\tau^{(c)}_{a|\mathcal{B}}$ for a particular example from PNAS, where the concept c is "nitrogen," the author a is Carl Nathan, MD, and the researcher has specified two immunology papers as his trusted set, $\mathcal{B} = \{10161, 9318\}$. For each paper $b \in \mathcal{B}$, we compute how much

the author a influenced b with respect to concept c. As our influence is now expressed from an *author* to an article, we treat all of an author's papers as a single unit. Specifically, we define the influence from author a to article b with respect to concept c, $AuthorInfluence_c(a \rightarrow b)$, as the probability there exists an active path in G_c from *any* article written by a to article b, where each edge is (independently) active with some probability $\theta^{(c)}_{x \rightarrow y}$. (Again, details on this definition and computation can be found in the full paper [22].)

In our example, we first look at how much Dr. Nathan's papers influence 10161 with respect to "nitrogen," and again from Dr. Nathan's papers to 9318. We now weigh these two influences by the prevalence of the word "nitrogen" in each paper b (as indicated by the pie chart in Figure 2.4), and define $\tau^{(c)}_{a|\mathcal{B}}$ to be the weighted sum of the two.

In order to incorporate trust into paper selection, we assume an author will trust a paper if and only if he trusts *at least one of its authors*. This intuition can be formalized by defining a modified notion of set influence, where the researcher's preferences towards the authors are directly taken into account:

$$Influence_c(q \leftrightarrow \mathcal{A}|\mathcal{B}) = 1 - \prod_{d \in \mathcal{A}} \left(1 - Influence_c(q \leftrightarrow d)\beta^{(c)}_d T^{(c)}_{d|\mathcal{B}}\right),$$

where $T^{(c)}_{d|\mathcal{B}} = 1 - \prod_{a \in authors(d)}(1 - \tau^{(c)}_{a|\mathcal{B}})$.

We can now define our *personalized* objective function as:

$$F_{\mathcal{Q}|\mathcal{B}}(\mathcal{A}) \quad = \quad \sum_{q \in \mathcal{Q}} \sum_{c \in \mathcal{C}} \gamma^{(c)}_q Influence_c(q \leftrightarrow \mathcal{A}|\mathcal{B}). \tag{2.11}$$

Maximizing $F_{\mathcal{Q}|\mathcal{B}}(\mathcal{A})$ subject to $|\mathcal{A}| \leq k$, for some budget of k papers, leads to a personalized set of papers tailored to someone who trusts $\mathcal{B}$. This function shares the same theoretical properties as Equation 2.10 and can be optimized efficiently in the same manner.

2.2.4 Evaluation

In order to truly evaluate our methodology we must solicit feedback from real scientific researchers. To this end, we conducted a user study involving sixteen subjects (all doctoral students in computer science or related fields).

We compare two variants of our algorithm–with and without incorporating trust preferences of the participant–with three representative alternative techniques: Google Scholar[4], Information Genealogy [56] (a hypothesis testing approach based on document text), and the Relational Topic Model [13] (a state-of-the-art topic model incorporating both text and citations to model latent themes in data). For each participant, we use each of these techniques to find related work for a previously written paper–that participant's *study paper*–thereby simulating a real research scenario. We define each query set $\mathcal{Q}$ to be the references of the corresponding study paper, and we ask each participant to list up to four trusted conferences or journals, which we use to define $\mathcal{B}$. For Google Scholar, we ask a coauthor of each participant to provide us with the keyword query he or she would use to find related work for their paper. The Relational Topic Model and Information Genealogy approaches each use the abstract of the study paper in their paper recommendation, which our algorithm does not require. Unlike many previous studies, each participant was asked

[4] `http://scholar.google.com`

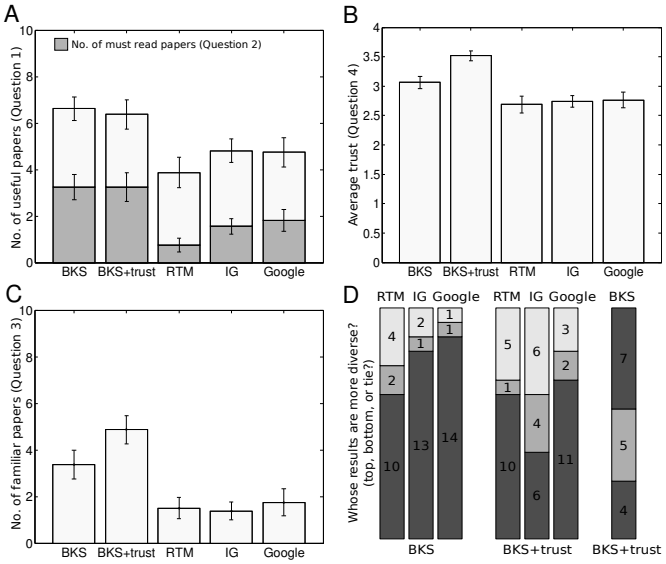


Figure 2.5: User study results comparing two variants of our algorithm, Beyond Keyword Search (BKS), with and without incorporating trust preferences, with the Relational Topic Model (RTM), Information Genealogy (IG) and Google Scholar. Values in bar plots (A), (B) and (C) are responses to the indicated study questions averaged over all sixteen participants, with error bars indicating one standard error. (D) shows how many participants (out of 16) found that our method produced more diverse results compared to the alternative techniques.

to evaluate all five comparison methods, rather than just a single technique. In total, 612 distinct papers were recommended using these five techniques across all sixteen participants. (The articles used in this study come from the ACM Digital Library[5].)

Each participant was presented with the recommended articles for his or her study paper in a double-blind fashion, masking the identity of the technique used to select each paper. Participants were asked to answer questions on the usefulness, novelty and trustworthiness of each paper with respect to their research. Additionally, participants were presented with entire result sets and asked to evaluate them in terms of diversity. Figure 2.5 shows the results of the study, from which we can glean the following main observations:

1. On average, users find the papers our algorithm selects to be more useful than those selected by the comparison techniques. The topic modeling approach performs especially poorly, with fewer than half of selected papers deemed useful.

2. Explicitly modeling the individual trust preferences of users leads to more trustworthy papers being selected. However, this comes at the expense of novelty in the selected articles, as researchers are more familiar with the work of authors they trust.

3. Our algorithm provides more diverse results than the comparison techniques, which is unsurprising, as our objective functions penalize redundancy.

[5] `http://portal.acm.org`

2.2.5 Summary

With this work, we demonstrated that the probabilistic max-cover framework can be successfully adapted to more complicated coverage settings (Q3). Moreover, we did so while allowing the user to specify his or her information need in a rich, natural manner (Q4). Finally, we demonstrated an approach for eliciting a user's trust preferences with limited effort on the user's part, and successfully incorporated this preferences into document selection.

Chapter 3

Proposed Work

3.1 Concept Granularity and Dynamics (Q1, Q2)

In our prior contributions, we made use of simple concept representations, including:

- Fine-grained concepts such as named entities, noun phrases or common words; and,
- Coarse-grained concepts such as topics from a latent Dirichlet allocation (LDA) topic model [6].

Despite their simplicity, they allowed us to make significant headway in addressing important information overload problems in the blogosphere and scientific literature domains, as described in the previous chapter. However, it also became clear that there was room for improvement that hinged on having more nuanced concept representations. In particular, we present the following four factors:

- In order to fully take advantage of the diminishing returns properties of our probabilistic max-cover framework, the concepts that we are covering need to represent fundamentally different ideas.
- From the standpoint of personalization, topics from a topic model are often too coarse to be single concepts, as each topic often represents many ideas, making it unlikely that a user has a single opinion about each concept when it comes to interest or trust.
- Moreover, as described in Chapter 1, the appropriate concept granularity for one person, say, a sports fan, might be completely different from that of someone else, say, a politico. Our previous work did not take this into account.
- In all cases, we need to be cognizant of the fact that document collections are often dynamic, and our representation should be able to seamlessly incorporate document additions and deletions over time. For instance, when we used coarse-grained concepts in TDN, we were forced to use a bipartite matching algorithm at each time epoch, to approximately match the topics in one epoch to the topics in the subsequent one. We desire to replace such a piecemeal approach by addressing dynamics in a more principled manner.

With this in mind, we have reason to believe that by *learning* an appropriate concept representation for a particular collection and for a particular user, we can do even better on the retrieval tasks we introduced previously. We plan on approaching this problem from two angles. First, we intend on leveraging the machinery of nonparametric Bayesian statistics to develop a probabilistic concept representation that falls in between fine-grained words and coarse-grained topics. We call this approach the *farmers market process*

(FMP),[1] following the culinary conventions of the field. Second, we plan on developing a new user interaction paradigm, that allows users to directly specify that particular results are redundant to them. We plan on using this *redundancy feedback* to inform the personalization of concept representations for specific users.

3.1.1 The Farmers Market Process

In this thesis, we assume that a document collection can be represented as a collection of concepts. Likewise, we can also assume that each concept is a collection of words (or other low-level features, such as named entities or noun phrases). However, *a priori*, we do not know how many concepts exist in a particular corpus, nor do we know how many words make up each concept. To model this setting, we turn to Bayesian nonparametric methods, which have been widely used in situations where there is uncertainty about the number of clusters or components in a model [46]. For example, the hierarchical Dirichlet process (HDP) is the natural nonparametric analog to the LDA topic model, where the number of topics is not known in advance [59].

Rather than having to specify the number of concepts in advance, we assume the existence of a countably infinite number of concepts, with only a finite number of them present in any fixed (finite) data set. Since we can view concepts as a particular kind of feature, the featural model of the Indian buffet process (IBP) and its underlying beta process is an appropriate starting point [26, 60].

Briefly, the IBP gets its name from an analogy to an Indian restaurant with an infinitely long buffet of dishes. The first customer starts at the front of the buffet and serves herself from a Poisson(α) number of dishes. Each subsequent customer serves himself or herself from the already sampled dishes with probability proportional to their popularity. The ith customer then samples an additional Poisson(α/i) new dishes.

If we view dishes as concepts and customers as documents, we end up with a representation where each document covers a small set of concepts. We note that this process provides a natural mechanism for dealing with a dynamic document collection, as new concepts can be introduced as new documents enter the corpus. Similarly, there is no need for any sort of bipartite matching, as was done in TDN, because users can define their preferences over concepts that persist over time. However, we have yet to define these concepts in terms of actual words. In order to do this, we represent each concept as a draw from a beta process, allowing for a concept to be defined by a set of words of unknown size. This *nested beta process* is in the spirit of the nested Dirichlet process of Rodríguez et al. [55], and was previously eluded to by Jordan [34].

To make this nesting clearer, in true nonparametric Bayesian fashion we propose the culinary analogy of a farmers market. A customer walks into a farmers market containing an infinite number of stalls. The first customer selects a Poisson(α) number of stalls to visit. The ith customer to enter the market goes to a previously visited stall k with probability proportional to the number of previous visitors to the stall. He or she then selects a Poisson(α/i) new stalls to visit. Each stall in the farmers market has an infinite number of items for sale, and again, the first visitor to a particular stall picks a Poisson(β) number of items to buy, while subsequent visitors to the same stall buy items proportional to their past popularity, and additional items in the usual Poisson fashion.

[1] Joint work with Emily Fox and Carlos Guestrin.

In this analogy, customers are documents, stalls are concepts, and items for sale correspond to words.[2] Recently, related work from Williamson et al. coupled an HDP with an IBP, resulting in what they refer to as a "focused topic model" (FTM) [63]. While in the future we envision incorporating a Dirichlet process component to our model to cluster together documents with similar concept representations, this particular work, while related, does not solve our concept granularity problem.

We believe that this nested beta process model and potential extensions of it (e.g., friends shopping together, signifying a Dirichlet process-style clustering of documents) can be effective in addressing concept granularity. Lingering questions remain, however, including:

- How do we perform efficient inference in this model for large document collections?
- Can this model be easily adapted to incorporate user feedback for personalized concept representations?

3.1.2 Redundancy Feedback and Personalized Concepts

Beyond modeling concepts at an appropriate *general* level of granularity, we wish to incorporate user feedback to tailor a concept representation to a *particular* user's beliefs. To use our running example, a sports fan might distinguish between the LeBron James and Wimbledon concepts, but someone who does not follow sports might find the differences between them minimal. In order to address such a scenario, we propose to 1) allow the user to provide direct feedback to indicate redundant documents, and 2) explore methods for *learning* a coverage-style objective function from this feedback.

Regarding the direct redundancy feedback, we can imagine an extremely simple user interaction, where upon seeing a result set of documents, the user can group the results into equivalence classes or clusters, where all the documents in a single cluster are considered by the user to be equivalent. By moving two documents into the same cluster, the user is telling us that while the algorithm believes there is some marginal benefit to be had for presenting one document to the user given that the other document has already been presented, the user disagrees.

We could imagine addressing this situation in two ways:

- We could keep the coverage function as is, but force a change in the underlying concept representation to reflect the feedback; or,
- We can keep the concept representation the same, but change the objective function itself by indicating that certain pairs of concepts may be referring to the same thing.

We find the second option to be more flexible, given that such an approach would apply to any concept representation we come up with, and so we plan on starting there. A particularly promising potential approach might involve recent work by Kulesza and Taskar [40] on structured determinantal point processes (SDPPs), which is a novel probabilistic model over sets that concisely captures the notion of *probabilistic mutual exclusion*, where similar objects (as determined by a kernel) are unlikely to appear together. We plan on investigating whether an SDPP (or similar model) learned in an online fashion from redundancy feedback data will lead to effectively personalized concepts.

[2] Moreover, making an analogy with the "super pixels" of computer vision, we refer to this type of concept as "super words."

3.2 Transferred Feedback for Trust (Q6)

In the introduction, we emphasized that while personalization is important for improving a user's utility from a retrieval system, it cannot come at a heavy cost in terms of time and effort spent by the user providing feedback. For this reason, in this thesis we explore methods of *transferring feedback* from one domain to another. Specifically, we wish to take information that a user has already provided that encodes some of his preferences, such as the people he follows on Twitter or the research articles he has previously cited, and use this information to inform the personalization of another task.

In Section 2.2, we described one such scenario where the trust preferences of a scientist were inferred from his or her citation history. We wish to extend this work to the blogosphere and online news.

3.2.1 Twitter-based Trust

Twitter[3] is a popular micro-blogging tool that allows users to post short messages, or "tweets," to their personal feed. As of mid-September 2010, Twitter claimed 175 million users who wrote 95 million tweets per day. Many of these users are celebrities, politicians and other public figures, as well as many major news organizations. For instance, individual columnists at the *New York Times* have their own Twitter accounts (e.g., http://twitter.com/nickkristof), which they use to post links to their articles, as well as other comments and documents that they may find interesting.

Users of Twitter interact with each other mainly through a "following" relationship. A user receives a feed containing all the tweets of the users he or she follows, and thus when one user decides to follow another, she is making some statement regarding her interests. Moreover, in addition to posting text, links and images, users can "retweet" the tweets of one of the users they follow, potentially indicating trust. Thus, as Twitter makes its data publicly available through an API, we feel that it provides a prime opportunity for demonstrating the utility of transferred feedback in a large-scale setting.

As a first question, we can consider how we might select blog posts as we did in TDN, but augmented with a Twitter-based trust model. Given that one is (or trusts) a particular Twitter user, what does this mean for selecting posts and articles? In TDN, as described in Chapter 2, we did not elicit or incorporate any trust preferences, and thus blog posts were judged solely on their content. While this may seem perfectly fine–egalitarian, even–we ended up with the situation where an arbitrary blog post from, say, a local newspaper in Guam, might be selected because it covers a particular politics story just as well as Politico or the *Washington Post*, which might be considered more authoritative sources for most users.

A simple change to the TDN model might ameliorate this problem, where we can define a new individual coverage function, $cover'_j(i)$, as the product of the old coverage function, $cover_j(i)$ and a trust multiplier, $\tau^{(j)}_{i|u}$, indicating how much Twitter user u trusts post j with respect to concept i. The resulting objective function would remain submodular, and the theoretical guarantees from TDN would still stand.

However, how do we actually compute this trust? We want to use some notion of transitivity as we did in Section 2.2, particularly based on the follower/followed and retweet relationships. If user A retweets a tweet from user B that covers some concept, and user B retweets some other tweet from a user C, that also covers that concept, then maybe user A might also trust user C when it comes to this concept. However, unlike research papers, whose citation graphs tend to be approximately acyclic, the follower/followed

[3] http://www.twitter.com

structure of Twitter certainly contains cycles. Thus, if considering a graphical model approach, we would need to work with an undirected model, as Bayesian networks require acyclicity. In particular, we might consider representing the Twitter graph as a conditional random field (CRF) [41], where random variables correspond to whether or not the current user trusts a particular feed with respect to a particular concept. These approaches will have to be investigated over the course of this thesis.

Another more subtle problem is that the vocabulary in Twitter is different than the vocabulary in news and blogs. Tweets are constrained to a maximum size of 140 characters, and thus contain many abbreviated words and phrases. The authors of a recent paper observed that about a quarter of the vocabulary in their Twitter data set was made up of words that do not appear in a standard English, French or Spanish dictionary, illustrating the uniqueness of the Twitter lexicon [21]. As such, in order for the trust values $\tau_{i|u}^{(j)}$ to be useful, we need the concepts in each of these two distinct corpora to be paired, such that a "Barack Obama" concept in the blogosphere (which might contain, e.g., the phrase "Mr. President") corresponds to the "Barack Obama" concept in Twitter (which may contain words like "prsdnt" or "prez"). While we hope to extend our nonparametric concept representation discussed in Section 3.1.1 to deal with such a situation, it is informative to first consider a simpler case where we use traditional topics as concepts, as we did in TDN+LDA.

One approach is to view the two corpora as being comprised of two languages, and hope to leverage prior work on multilingual topic modeling [8, 45]. Unfortunately, this work is not an ideal match for this situation as it tends to assume paired documents (i.e., the same document in multiple languages), such as a French Wikipedia article and an English Wikipedia article about the same subject, or barring that, some sort of dictionary providing translations of words.

A promising alternative is to leverage a model similar to the one proposed by Eisenstein et al. [21], which seeks to model regional variation in language by defining "regional topics" as a corrupted version of some base topic. For example, there might be a base sports topic that has regional variants in New England that contains high weight for words like "Celtics," and Southern California which emphasizes words like "Lakers." A key insight is that if we view Twitter and the blogosphere as if they were two distinct geographical regions, we feel that we might be able to successfully model such variation.

3.3 Modeling Personalization Dynamics (Q5, Q7)

3.3.1 Changing User Interests

While we argue that personalization is often desirable, there are distinct cases where a user's opinions or interests change over time, leading to outdated recommendations. For example, consider a hypothetical user who usually does not care about tennis, with the important exception of the two weeks of Wimbledon, when he follows tennis intently. The ideal behavior for a retrieval system for such a user would be to present him with tennis articles during Wimbledon, but quickly back off directly afterward.

Our online learning approach in TDN did not deal with this situation, rather assuming static user interest in the various concepts. In fact, in standard fashion, we decrease our learning rate as the number of epochs of personalization increases, as we expect to be closer to convergence to a user's true (static) preferences. As part of this thesis work, we wish to study methods from the online learning community–particularly those that deal with concept drift–for addressing this problem. Perhaps we can learn that users tend to exhibit

more varied interest over time in some classes of concepts than others, and by modeling such variance, we might achieve more realistic personalization behavior.

3.3.2 Overpersonalization

Personalization is a ubiquitous feature in today's Web experience, from search to online shopping to news feeds. An interesting question to address is, what happens if a user experiences "too much" personalization? In fact, is there such a thing as overpersonalization? While previous work has shown the existence of such an "echo chamber" effect [30, 57], we are interested in seeing if we can observe and measure this phenomenon in large-scale online retrieval systems.

For instance, perhaps by looking at a large corpus of research papers, we can study how the success of scientists who cite papers from a broad range of fields differs from those who cite narrowly. We can also imagine running behavioral experiments, potentially collaborating with the Psychology Department and the Human-Computer Interaction Institute.

3.4 Structured Input and Output (Q3, Q4)

In Section 2.2, we argued that when it came to search for related work for a specific research problem, it was often more natural to specify the query as a set of papers rather than as a string of keywords. In the same vein, it is interesting to consider richer representations of a user's information needs. In particular,

1. When specifying a set of research papers, a user might know that a subset of them are all intimately related, e.g., a collection of topic modeling papers. Thus, rather than specifying a flat set of query documents, the user can be more informative by specifying clusters or hierarchies of documents.
2. Sometimes, a researcher is interested in a paper only for a particular section. For instance, continuing with the topic model example, a researcher might want to find work that is related to the variational inference section of the LDA paper [6], but does not care much for other details particular to the model.

We intend on addressing these rich models of query behavior over the course of this thesis.

On the other end of the user interaction spectrum, we consider the situation where a simple ranked list of search results is not the ideal output for a particular retrieval system. Specifically, if a researcher's information need is sufficiently broad, she may not find a small list of papers useful. For example, Figure 3.1 shows a screenshot of the top Google Scholar search results for the query "machine learning." While this collection includes very heavily cited papers and books, including some seminal work (e.g., Quinlan's paper on decision trees), a scientist from a different discipline wishing to become acclimated to machine learning is unlikely to find these results particularly illuminating. We believe the reason for this is twofold:

- The list shows a simple selection of articles, with no context as to how they are situated within the field of machine learning. For example, it would be nice to know that Q-learning falls within a subfield called "reinforcement learning."
- The presented results are skewed towards old, seminal articles, with no reference to the current state of the art.

As such, we propose to investigate methods for *mapping the research landscape* of a particular area. For instance, we can imagine the output of a broad query like "machine learning" consisting of clusters corresponding to subareas like "reinforcement learning," "online learning," "transfer learning," "probabilistic graphical models," "maximum margin methods," and the like, with just one or two representative papers highlighted in each subarea. These representative papers might be identified as being seminal, state-of-the-art or tutorials.

Here, we seek to build upon prior work in the information retrieval community on both subtopic retrieval [65] and the clustering of search results (e.g., `www.clusty.com`) by *simultaneously* discovering clusters that cover a particular area as well as appropriate representative papers belonging to each cluster. To this end, we have performed initial experiments focusing on using a submodular coverage function for clustering, following Narasimhan et al. [48]; we wish to investigate this further over the course of the thesis.

3.5 Expected Contributions

At the conclusion of this thesis work, we expect to have made the following contributions:

- A method for automatically learning concept representations for a dynamic document collection.
- A method for personalizing the concepts of a document collection given limited user feedback.
- A general framework for solving information retrieval tasks using submodular coverage of concepts.
- An algorithm for learning user preferences for concepts from limited feedback data, where user interests are assumed to change with time.
- An approach for incorporating a user's trust preferences into retrieval tasks via transferred feedback.
- A study of overpersonalization behavior on the Web.
- Methodology for incorporating structured input and output into retrieval models.

Furthermore, to demonstrate these contributions, we plan on developing two systems that will bring them together in a real-world setting:

- A personalized blog post recommendation system, selecting posts that cover the main stories of the day, but tailored to the interests and trust preferences of a user.
- A rich search engine for research literature, allowing for both structured input and output, and incorporating trust preferences of the user.

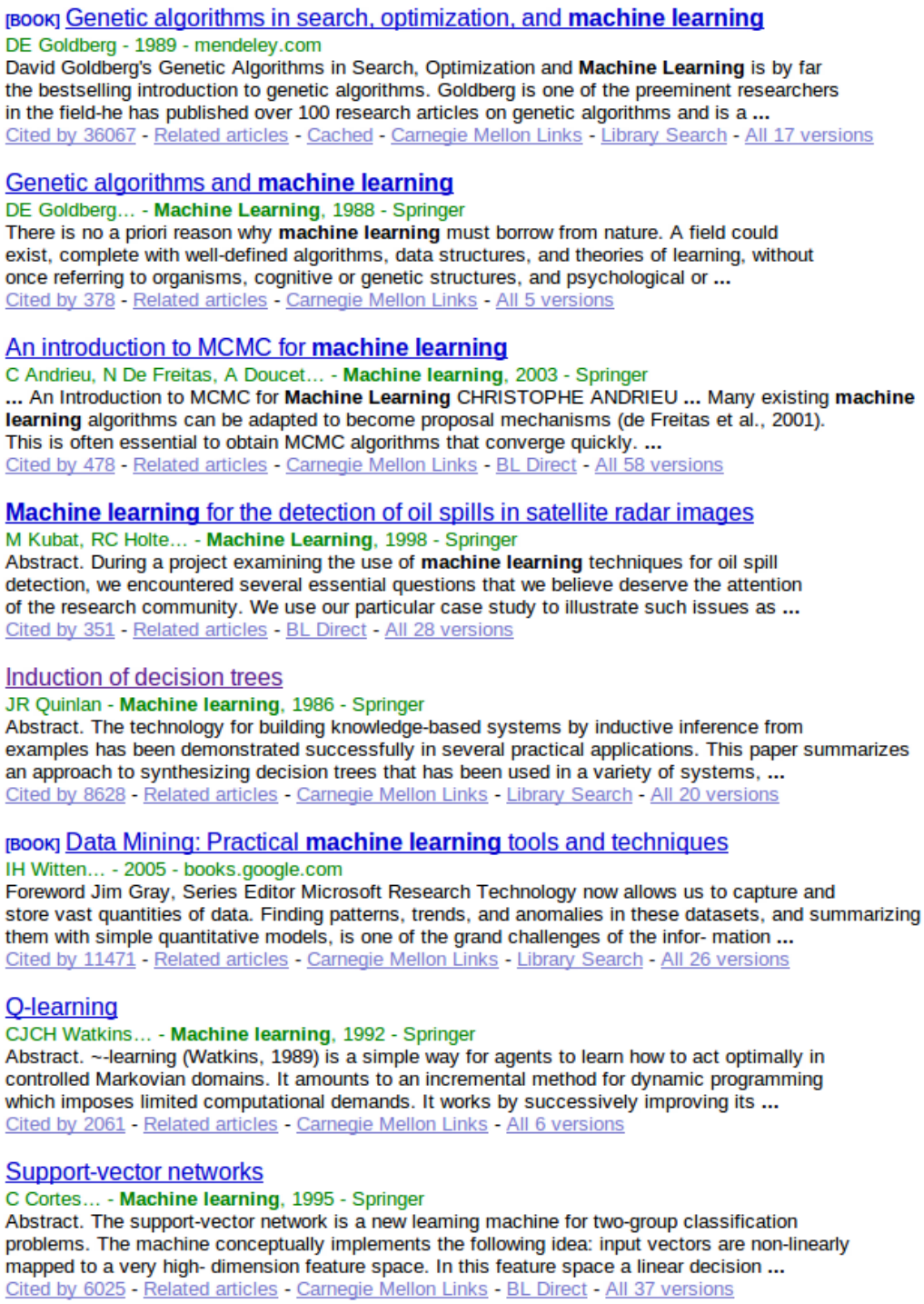
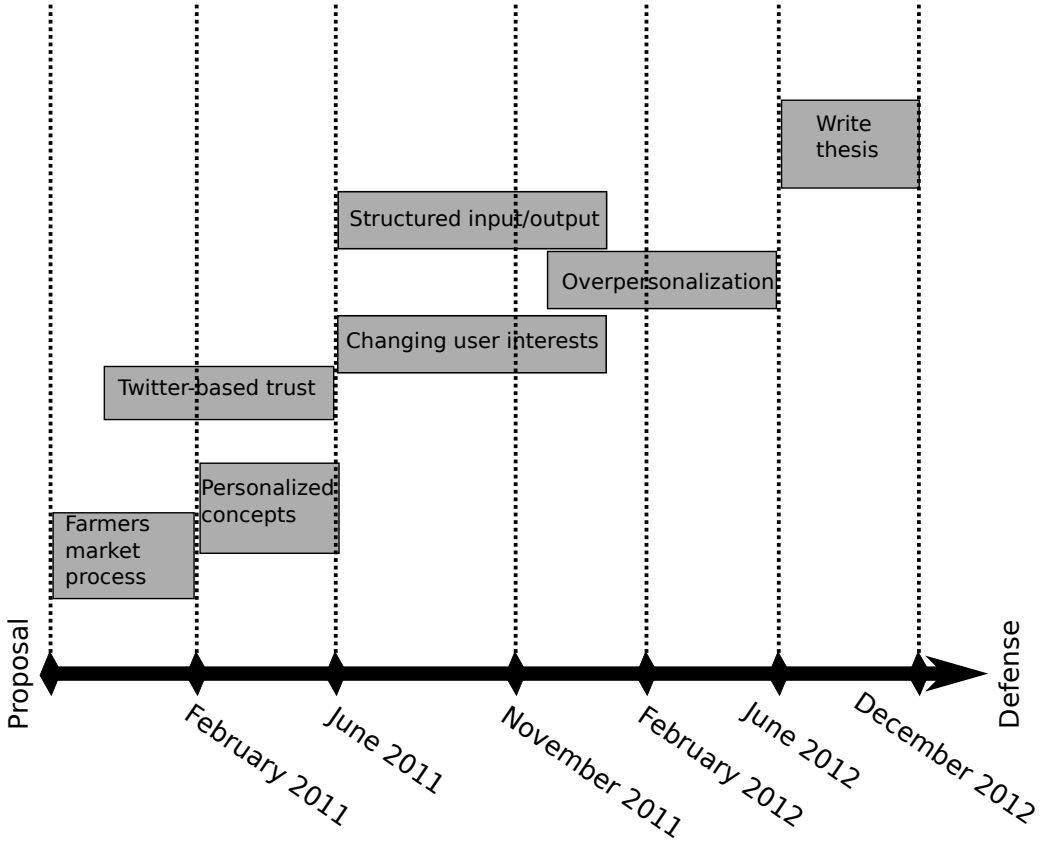


Figure 3.1: A screenshot of the top results returned by Google Scholar on November 17, 2010, for the query "machine learning."

Chapter 4

Detailed Timeline



Bibliography

[1] R. Adler, J. Ewing, and P. Taylor. Citation statistics. *Statistical Science*, 24:1–14, 2009. 1.3

[2] R. Andersen, C. Borgs, J. Chayes, U. Feige, A. Flaxman, A. Kalai, V. Mirrokni, and M. Tennenholts. Trust-based recommendation systems: an axiomatic approach. In *WWW*, 2008. 1.3

[3] M. Bates. The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13:407–424, 1989. 1.3

[4] D. M. Blei and J. Lafferty. Dynamic topic models. In *ICML*, 2006. 1.1

[5] D. M. Blei and J. Lafferty. Topic models. In A. Srivastava and M. Sahami, editors, *Text Mining: Theory and Applications*. Taylor and Francis, 2009. 1.1

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. 3.1, 2

[7] K. Bollacker, S. Lawrence, and C. L. Giles. Discovering relevant scientific literature on the Web. *IEEE Intelligent Systems and their Applications*, 15:42–47, 2000. 1.3

[8] J. Boyd-Graber and D. M. Blei. Multilingual topic models for unaligned text. In *UAI*, 2009. 3.2.1

[9] R. J. Brachman and H. J. Levesque. *Knowledge Representation and Reasoning*. Morgan Kaufmann, 2004. 1.1

[10] K. R. Canini, L. Shi, and T. L. Griffiths. Online learning of topics with latent dirichlet allocation. In *AISTATS*, 2009. 1.1, 3

[11] J. Carbonell and J. Goldstein. The use of MMR, diversity-based re-ranking for reordering documents and producing summaries. In *SIGIR*, 1998. 1.2

[12] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006. 2.1.2

[13] J. Chang and D. M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, 4:124–150, 2010. 2.2.4

[14] H. Chen and D. Karger. Less is more. In *SIGIR*, 2006. 1.2

[15] P. Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with Google. *Journal of Informetrics*, 1:8–15, 2007. 1.3

[16] A. S. Das, M. Datar, A. Garg, and S. Rajaram. Google News personalization: scalable online collaborative filtering. In *WWW*, 2007. 1.3

[17] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990. 1.1

[18] L. Dey, A. C. Rastogi, and S. Kumar. Generating concept ontologies through text mining. In *International Conference on Web Intelligence*, 2006. 1.1

[19] I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clusters. *Machine Learning*, 42(1):143–175, 2000. 1.1

[20] D. Diderot. In D. Diderot and J. d'Alembert, editors, *Encyclopedia, or a systematic dictionary of the sciences, arts and crafts*, Paris, 1755. Briasson, David, Le Breton, and Durand. (tr. from French). 1

[21] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *EMNLP*, 2010. 3.2.1

[22] K. El-Arini and C. Guestrin. Beyond keyword search: Discovering relevant scientific literature. In *NIPS Workshop on Networks Across Disciplines: Theory and Applications*, 2010. 1.2, 2.2, 2.2.2, 2.2.3

[23] K. El-Arini, G. Veda, D. Shahaf, and C. Guestrin. Turning down the noise in the blogosphere. In *KDD*, 2009. 1.2, 2.1, 2.1.4

[24] Y. Freund and R. E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 1999. 2.1.2, 2.1.2

[25] E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178:471–479, 1972. 1.3

[26] T. L. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *NIPS*, 2006. 3.1.1

[27] T. R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993. 1.1

[28] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *WWW*, 2004. 1.3

[29] J. A. Hampton. Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 18:441–461, 1979. 1.1

[30] C. Harper. The Daily Me. *American Journalism Review*, April 1997. 3.3.2

[31] Z. Harris. Distributional structure. *Word*, 10(2/3):146–162, 1954. 1.1

[32] B. Hayes-Roth and F. Hayes-Roth. Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning and Verbal Behavior*, 16:321–328, 1977. 1.1

[33] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences USA*, 102:16569–16572, 2005. 1.3

[34] M. I. Jordan. Hierarchical models, nested models and completely random measures. In M.-H. Chen, D. Dey, P. Müller, D. Sun, and K. Ye, editors, *Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger*. Springer, 2010. 3.1.1

[35] S. Khuller, A. Moss, and J. Naor. The budgeted maximum coverage problem. *Information Processing Letters*, 1999. 2.1.1

[36] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46:604–632, 1999. 1.3

[37] J. Z. Kolter and M. A. Maloof. Dynamic weighted majority: An ensemble method for drifting concepts. *Journal of Machine Learning Research*, 8:2755–2790, 2007. 1.3

[38] A. Krause and C. Guestrin. Optimizing sensing from water to web. *IEEE Computer Magazine*, August 2009. 1.2

[39] A. Krause, H. B. McMahan, C. Guestrin, and A. Gupta. Robust submodular observation selection. *Journal of Machine Learning Research*, 9:2761–2801, 2008. 1.2

[40] A. Kulesza and B. Taskar. Structured determinantal point processes. In *NIPS*, 2010. 3.1.2

[41] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001. 3.2.1

[42] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *KDD*, 2007. 1.2, 2.1.1

[43] G. Linden, B. Smith, and J. York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7, 2003. 1.3

[44] S. M. McNee, I. Albert, D. Cosley, P. Gopalkrishnan, S. K. Lam, A. M. Rashid, J. A. Konstan, and J. Riedl. On the recommending of citations for research papers. In *CSCW*, 2002. 1.3

[45] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum. Polylingual topic models. In *EMNLP*, 2009. 3.2.1

[46] P. Müller and F. A. Quintana. Nonparametric bayesian data analysis. *Statistical Science*, 19(1):95–110, 2004. 3.1.1

[47] G. L. Murphy. *The Big Book of Concepts*. MIT Press, 2002. 1.1

[48] M. Narasimhan, N. Jojic, and J. Bilmes. Q-Clustering. In *NIPS*, 2005. 3.4

[49] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of the approximations for maximizing submodular set functions. *Mathematical Programming*, 14:265–294, 1978. 2.1.1

[50] C. Olston and E. H. Chi. Scenttrails: Integrating browsing and searching on the Web. *ACM Transactions on Computer-Human Interaction*, 10:177–197, 2003. 1.3

[51] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the Web. Technical report, Stanford University InfoLab, 1999. 1.3

[52] S. Pandit and C. Olston. Navigation-aided retrieval. In *WWW*, 2007. 1.3

[53] A. Plangprasopchok, K. Lerman, and L. Getoor. Growing a tree in the forest: constructing folksonomies by integrating structured metadata. In *KDD*, 2010. 1.1

[54] F. Radicchi, S. Fortunato, B. Markines, and A. Vespignani. Diffusion of scientific credits and the ranking of scientists. *Physical Review E*, 80:056103, 2009. 1.3

[55] A. Rodríguez, D. B. Dunson, and A. E. Gelfand. The nested Dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1154, 2008. 3.1.1

[56] B. Shaparenko and T. Joachims. Information genealogy: uncovering the flow of ideas in non-hyperlinked document databases. In *KDD*, 2007. 2.2.4

[57] C. R. Sunstein. *Going to Extremes: How Like Minds Unite and Divide*. Oxford University Press, 2009. 1.3, 3.3.2

[58] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *SIGIR*, 2005. 1.3

[59] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. 3.1.1

[60] R. Thibaux and M. I. Jordan. Hierarchical beta processes and the Indian buffet process. In *AISTATS*, 2007. 3.1.1

[61] R. Torres, S. M. McNee, M. Abel, J. A. Konstan, and J. Riedl. Enhancing digital libraries with TechLens+. In *JCDL*, 2004. 1.3

[62] X. Wang and A. McCallum. Topics over time: A non-markov continuous-time model of topical trends. In *KDD*, 2006. 1.1

[63] S. Williamson, C. Wang, K. A. Heller, and D. M. Blei. The IBP compound Dirichlet process and its application to focused topic modeling. In *ICML*, 2010. 3.1.1

[64] Y. Yue and T. Joachims. Predicting diverse subsets using structural SVMs. In *ICML*, 2008. 1.2

[65] C. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR*, 2003. 1.2, 3.4