Understanding Document Aboutness
Step Two: Identifying Interesting Things

Michael Gamon MGAMON@MICROSOFT.COM
Microsoft Research
Redmond, WA, USA

Arjun Mukherjee* ARJUN4787@GMAIL.COM
Department of Computer Science & Engineering
University of Illinois at Chicago, Chicago, IL 60607, USA

Patrick Pantel PPANTEL@MICROSOFT.COM
Microsoft Research
Redmond, WA, USA

Abstract

We define the notion of an interesting nugget in a document. Such nuggets attract a user's attention and lead them to explore more information around that nugget. In order to measure and model interestingness, we look at browsing sessions within Wikipedia and we build a data set of transitions (clickthrough) from a source Wikipedia page to a destination Wikipedia page through anchor clicks. We investigate factors that influence the probability of a click on an anchor in a Wikipedia page. We propose a topic modeling approach which jointly models the contents of the source and destination pages. We then use the estimated posterior on latent variables as features, along with page structure and user metadata features to build a model of interestingness. Finally, we evaluate this model using different feature sets and we demonstrate the model's effectiveness at predicting interesting nuggets. Experimental results show that the latent semantic features are effective in predicting interestingness and can outperform baseline features.

1. Introduction

While an enormous amount of research has been done on information-seeking or, more recently, sense-making in a decision process (Evans & Chi, 2001), there has been much less emphasis on what we call *interestingness*: the – sometimes serendipitous – encounter of a concept, entity, or fact (henceforth dubbed *nugget*) during content consumption that leads the user to investigate a particular *interesting* nugget further. Sometimes this behavior leads to the exploration of a single nugget, sometimes the user is enticed into a lengthy chain of following one interesting nugget to another and so on. For example, the user might consume content about an actress and her recent and successful TV series. They might then click on a link to a biography of the actress, where they might find links to other TV shows that now pique their interest.

Being able to predict what is *interesting* in the context of some specific content is useful in a number of ways. We can enrich the content consumption experience for a user by providing augmentation of document content: We can automatically turn the interesting nuggets into anchors which link to a URL with relevant information. This kind of augmentation invites the user to explore interesting avenues

* This work was conducted at Microsoft Research.

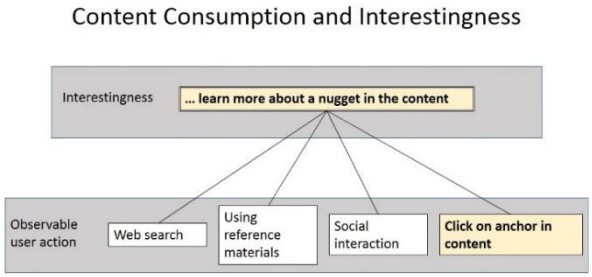

Figure 1: Content Consumption and Interestingness.

from the current document without cluttering the content with uninteresting anchors. From a purely text content based perspective, knowing interesting and uninteresting entities in a document can improve concept decomposition for large scale text clustering, building improved concept dictionaries/associations (Okumura & Hovy, 1994) and concept ontologies (Maedche & Staab, 2000). Within a web context, characterization of interesting concepts/anchors in a page has promising applications in concept clustering/representation (Huang et al., 2009), building concept hierarchies (Sanderson & Croft, 1999), latent concept expansion (Metzler & Croft, 2007), and tagged web clustering (Ramage et al., 2009). Further, since discovering interesting anchors is an instance of web page annotation/augmentation with additional metadata, it holds promise for various downstream tasks like improved web search relevance (Bao et al., 2007), topic-link based document retrieval using language models (Chang et al., 2009), and improved information presentation for browsing (McKeown et al., 2002) where augmentation has been shown to be very successful.

Upon closer reflection it becomes clear that this notion of *interestingness* is not a static concept. What is perceived as *interesting* in the context of some specific content depends on the current content (source), the content that could be provided for a nugget (destination) and the user. An item may be interesting to most readers in one document, but not in another. Consider our previous example, a popular actress who is mentioned in a document about her latest successful TV series. Users might find it interesting to follow up on her in this context, while the same actor mentioned in a document about actors who have worked for a specific film studio in the 1990s might elicit less of an interest from a general audience. Similarly, content about the actress and her acting career might be of higher interest than content about her clothing line.

Figure 1 illustrates our view of content consumption and interestingness. While consuming content, the user is interested in learning more about some nugget in the current context. This interest leads to an observable user action, which can take many different forms. For example, the user might issue a query to a search engine, consult a printed reference such as a dictionary, consult with friends over a social network, or (in web content consumption) click on an anchor on the current page. All these observable user actions can be used as indication that the user shows interest in the given nugget.

A central concept for our modeling of interestingness is a *transition t(S, D)* from a source page *S* to a destination page *D*, mediated by an anchor. We take these transitions as driven by the interestingness of the anchor in *S* that links to *D*.

The main contributions of this paper are:

- We introduce a notion of interestingness that is grounded in observable behavior during content consumption.
- We propose a probabilistic model of interestingness, which captures the latent aspects that drive a user to be interested in browsing from one document to another, by jointly combining the latent semantics of the source and destination documents with the intent of the transition.
- We pose the problem of predicting interesting browsing transitions as a discriminative model combining evidence from our model of interestingness, contextual features in the source document, and geospatial and time features. We train our models on millions of real-world browsing events.
- We empirically show that our model is effective and that the latent semantic features contribute significantly on the task of predicting the most interesting links to a user in a web browsing scenario.

2. Related Work

Our work focuses on building models of interestingness, based on contextual information from both the currently consumed document and the document that could be linked to from an interesting nugget. There are several areas of related work which we summarize below under specific research heads.

2.1. Click Prediction

Click modeling aims to interpret the users' click data in order to predict their clicking behavior. Click prediction models are usually built using click-through logs from search engines for applications like Web search ranking, ad click-through rate (CTR) prediction and personalized click recommendation models.

One of the main foci of click prediction models is query based click prediction which aims at computing the probability that a given document in a search-result page is clicked on after a user enters some query. The main technique is to learn the user-perceived relevance for query-document pairs. This has attracted significant amount of attention and several attempts have been made to improve the overall search results (Agichtein et al., 2006; Guo et al., 2009; Joachims, 2002; Joachims et al., 2005). In most click prediction models, dwell time is used as an intrinsic relevance signal (Buscher et al., 2009; Kelly & Belkin, 2004; Morita & Shinoda, 1994; White & Kelly, 2006). Yet there is other work that has investigated the position bias problem (Granka et al., 2004), revisiting behaviors (Xu et al., 2012), post-click behaviors (Guo & Agichtein, 2012), and freshness for news search (Wang et al., 2012).

A key distinction from the approaches above is that our models are capable of learning notions of interestingness from document (semantic) content and browser page transition click logs.

2.2. Online Advertising and Sponsored Search

Sponsored search (i.e., search results with recommended ads) is another area where click prediction models have been shown to be effective. Ads are ranked according to their likelihood of being relevant

to the user and their likelihood of generating high revenue, which are highly correlated with the probability of a user click. This makes accurate click prediction crucial to sponsored search. Several approaches have been proposed which include ad click prediction using historical Click Through Ratio (CTR) (Graepel et al., 2010; Craswell et al., 2008) and references therein; exploiting the semantic relevance of query and ad content (Hillard et al., 2011; Richardson et al., 2007); exploiting mutual influence between ads (Ghosh & Mahdian, 2008; Xu et al., 2010); relation click prediction based on multiple ads on a page (Xiong et al. 2012), understanding the positional bias in sponsored search (Chen & Yan, 2012)0, using multimedia features in ad click prediction (Cheng et al., 2012), etc. In (McMahan et al., 2013), the authors perform case studies on large scale evaluation of ad click prediction and shed light on many practical aspects like efficient memory usage, calibration techniques, and feature management.

There have also been studies which measure the effectiveness of online targeted advertising (Farahat & Bailey, 2012) and dynamics of advertiser bidding (Xu et al., 2013). Also related are personalized click models e.g., (Shen et al., 2012) which tend to use user CTR and various browsing metadata (e.g., demographic history (Cheng & Cantú-Paz, 2010)) to improve personalized and sponsored search.

Although ad click prediction is related to our task setting of predicting interesting anchors on a page, the methods used in ad click prediction are not directly applicable as they rely on ad query and query augmented metadata. Our models are grounded in browsing behavior (Figure 1) and hence can only rely on document content and structure.

2.3. Contextual Advertising

Contextual advertising (Broder et al., 2007) places ads within the content of generic third party web pages. There is usually some commercial intermediary (ad-network) which is in charge of optimizing the ad selection with the dual goal of improving revenue and user experience. Studies have shown that its success is closely related to accurate click prediction (Chatterjee et al., 2003). Ad representation with word/phrase vectors have been shown to work well (Broder et al., 2007; Lacerda et al., 2006). Extensions include models which combine click feedback (Chakrabarti et al., 2008), forecasting ad-impressions (Wang et al., 2009), etc. However all models in this thread primarily rely on semantic match of the content page and ad which is different from our task setting of predicting interesting browsing transitions from a page.

2.4. Use of Probabilistic Models

The majority of the research summarized above employs probabilistic models owing to the very nature of click prediction's reliance on historical CTR. Our semantic model is built over LDA (Blei et al., 2003) and has resemblances to Link-LDA (Erosheva et al., 2004) and Comment-LDA (Yano et al., 2009) models. However, those models are tailored for blogs and associated comment discussions which is very different from our source to destination transitions of user browsing from web browser logs. Also related are the approach of (Gao et al., 2011) which employs statistical machine translation to learn semantic translation of queries to document titles. Guo et al. (2009) used probabilistic models for discovering entity classes from query logs and Lin et al. (2012) studied latent intents in entity centric search. The above models apply to individual pages/queries and associated metadata and do not model the joint topical/concept mappings that are involved in source to destination page transitions which is the focus of our work.

2.5. Predicting Popular Content

Modeling interestingness is also related to work which has focused on predicting popular content in the Web. This work mainly focuses on news content popularity and prediction (Lerman & Hogg, 2010; Szabo & Huberman, 2010), and popular content in the Web (Bandari et al., 2012). Our models differ in the fact that we strive to predict what a user is likely to be interested on a page when consuming content. Our models do not rely on prior browsing history from click logs, since we strive to predict interestingness in situations where we have little or no history. Instead, our models make use of learned higher level semantic concept mappings learned from article contents.

3. Interestingness

As we have outlined in the introduction, *interestingness* manifests itself in observable click behavior; this forms the signal which we utilize as the target for our models.

We treat the anchors on a page as the set of candidate nuggets. Utilizing the observed click behavior as a proxy signal for *interestingness,* we can build models of interestingness to approximate the nugget/anchor click behavior by using features from the current content, the destination content, and the user. Note that even though we focus on browsing behavior as our signal, this does not limit the application of our models to the web alone. Since we use general document level and semantic features in our models, hence the models can also be applied to documents in general, even though they are trained on web browsing data.

Our data consists of randomly sampled browsing transitions (clicks on anchors) from a source page to a destination page. In our study, we focus specifically on transitions within Wikipedia, for a number of reasons. First, Wikipedia is a popular and much consulted resource, hence there is ample opportunity to observe how users click on anchors in one Wikipedia page and land on another while consuming content, alleviating problems of sparsity with generic web content. Second, Wikipedia is highly structured, allowing us to reliably extract content and anchors.

To obtain a better understanding of the problem, we performed a number of manual studies on a small random sample of 200 transitions from our data set and we also performed several simple prediction experiments to assess the predictiveness of several factors. Below, we summarize some of our observations and results of manual annotation experiments.

3.1. Only Few Things on a Page are Interesting

The average number of anchors on a Wikipedia page is 79. Out of these many anchors, only very few are actually followed by users. For example, the Wiki article on the popular TV series "The Big Bang Theory" leads to transitions to the pages of the actors in that series for the vast majority of transitions. 90% of the transitions are to the four pages for Kaley Cuoco, Jim Parsons, Kevin Sussman and Johnny Galecki.

3.1.1. The Semantics of Source and Destination Page is Important

Not surprisingly, we found that our collection of transitions within Wikipedia is governed by general popular interests. We manually determined the entity type of the Wikipedia articles in our sample (according to schema.org classes). 49% of all source urls in our data sample are of the "Creative Work"

category, reflecting the strong popular interest of movies (37%), actors (22%), artists (18%), and television series (8%). The next three most prominent categories are "Organization" (12%), "Person" (11%) and "Place" (6%).

We also observed that transitions are influenced by the category of the destination page. If the source article category is "CreativeWork.Movie", the most frequent destination categories are "Actor" (63%) and "Character" (13%). For the source category "CreativeWork.TVSeries", the "Actor" destination accounts for 86%, and the "Artist" category for the remaining 14% of transitions. "CreativeWork.Actor" favors the destination article category "Movie" (45%) and "Actor" (26%), while a source category "CreativeWork.Artist" favors destinations of category "Artist" (29%), "Movie" (17%) and "MusicRecording" (18%).

This suggests strongly that the semantics of both source and destination pages play an important role in what users find interesting.

3.1.2. The User Plays a Role

We were also interested in the question how the user and time information play into interestingness. For example, it is a reasonable hypothesis that users from very different geographic (and hence cultural) backgrounds might show different patterns in what they find interesting. Similarly, it is not unreasonable to assume that browsing behavior during weekends might be different from the behavior on week days. To investigate these hypotheses, we trained a series of simple Naïve Bayes models to predict the most interesting anchor on a page. Naïve Bayes models took a given feature dimension into consideration by modeling $p(d|s, f)$, where d, s, f denote the destination page, source page, and feature dimension. Our Null Model (**NM**) simply used the prior distribution $p(d|s)$ observed from the training data. We found that adding the user city, and user country feature dimensions in our Naïve Bayes model, respectively led to small but significant (p<0.05) improvements in prediction accuracy over the null model. Adding the feature dimension time of day (in 24h intervals) and the day of week to the Naïve Bayes model also led to statistically significant improvements over the null model.

3.1.3. The Structure of the Source Page Plays a Role

It is well known that the position of a link on a page influences user click behavior: links that are higher on a page or in more prominent position tend to attract more clicks. We verified that this effect also holds for our prediction task. When anchor position was added to the Naïve Bayes model as a feature, we observed a significant improvement over the null model baseline.

4. Predicting Interestingness

The general objective of identifying interesting nuggets on a page consists of two parts: (1) a set of nugget candidates for interestingness is identified; and (2) interestingness scores are assigned to each candidate. In this paper, we restrict our scope to the second task, i.e., the assignment of an interestingness score to a candidate nugget, where a candidate nugget is an anchor. We observe user clicks on anchors in a Wikipedia page, the set of candidate nuggets for us is the set of anchor texts on a source page. We believe that this narrower scope is appropriate in order to understand the factors that enter into what is perceived as interesting by a content consumer. On the other hand, we also believe that once we have

gained an initial understanding of the interestingness scoring problem, there are many intriguing research opportunities in identifying nugget candidates automatically.

We distinguish two tasks settings where interestingness scores can be used, and we evaluate them separately. The first task setting is to propose *k* anchors on a page that the user will find interesting (*highlighting* task). The second, and more difficult, setting is to predict which anchor the user is going to click on next (*click prediction* task).

4.1. Data Set

For our investigation we utilize two data sources. First, we use the full dump of English Wikipedia[2] containing about 4.1 million unique article pages as our universe of content. Second, we use one month of web browser log data joined against the Wikipedia data to capture instances where a user transitioned from one Wikipedia page to another by clicking on one of the anchors of the article. The browser log data provide us with metadata for each such transition from a source (s) Wikipedia page to a destination (d) article, including user time, location, and dwell time. We refer to a pair (s, d) of urls where a user clicked on an anchor in S and continued browsing the content of D as a *transition*. Our data set includes millions of transitions between many of the Wikipedia pages.

Using uniform sampling, we split our data into 20% development set, 60% training set and 20% test set. We further divide the test set into a HEAD, TORSO, and TAIL set using inverse CDF sampling on a traffic-weighted set in order to be able to evaluate the performance of our models on these three segments separately. Both HEAD and TAIL set account for 20% of the (source) traffic, while the TORSO set accounts for the remaining 60%.

4.2. The Prediction Task

Let U be the set of all Wikipedia articles and A the set of all anchors in U, respectively. Let $A_u \subset A$ be the set of anchors in $u \in U$. We formally define the interestingness task as learning the function:

$$\boldsymbol{\sigma}: \boldsymbol{U} \times \boldsymbol{A} \rightarrow \mathbb{R} \tag{1}$$

where $\sigma\ u, a\ $ reflects the interestingness of a in *u* [3].

For our learning algorithm we use boosted decision trees (Friedman, 1999) to learn a regression model. Hyperparameters such as number of iterations, learning rate, minimum instances in leaf nodes, and the maximum number of leaves are tuned using 3-fold cross-validation on the training data.

The observed interestingness scores which our model regresses to are derived from user clicks on anchors in a source page s leading to a destination page d. Specifically, the regression target is $p(d|s)$, where this probability is estimated using the aggregated click counts on d in s and the total number of clicks on any anchor in s.

Each anchor/document pair is represented as a vector of features, where the features fall into four basic categories:

[2] Our English Wikipedia corpus consists of the May 3, 2013 dump available at http://dumps.wikimedia.org.
[3] We fix $\sigma\ u, a\ = 0$ for all $a \notin A_u$.

1. Geotemporal features (**GeoTemp**): country, city, postal code, region, state, timezone, time of click, day of click
2. Anchor features (**Anc**): position of the anchor in the document, frequency of the anchor, density of the anchor in the paragraph where it was clicked, whether the anchor text matches the title of the destination page
3. Semantic features based on manual categorization by the Wikipedia editors (**Wiki**)
4. Semantic features derived from an unsupervised joint topic model of source and destination pages (as described in detail in Section 5)

5. The Semantics of Interestingness

5.1. Motivation and Model Overview

Our goal in this paper is to learn a model of interestingness that is capable of scoring an anchor text according to the probability that this anchor text is of interest to a user when he is consuming/browsing the article content.

As we have seen in Section 3.1.1, the semantics of the source and destination page play a role in whether an anchor is perceived as interesting. This reflects the intuition that notions of human interests are usually encoded in higher level concepts or semantic spaces which are influenced by article contents. For example, recall from Section 3.1.1 that when people are consuming a movie page, they are likely to be interested in the lead actor/actresses/director associated with the movie. However, most movie pages in Wikipedia almost always have other related information like shooting venue, release budget, sales, critics, etc. which are rarely clicked showing a lower interest/learning need of users. Here we can find an interest mapping of movie $\rightarrow$ artist where movie and artist are some higher level semantic abstractions of all movie and actor pages.

In Wikipedia, editors assign categories to articles, so we can get semantic information "for free" in order to measure the influence of the content semantics of source and destination page on interestingness. As we will see in Section 6.1.1, this information indeed influences interestingness considerably. Knowing that semantics does play a big role motivates us to build an unsupervised semantic model of source and destination pages that does not rely on manual annotation. Such a model can then serve the purpose of providing important semantic signals for interestingness in a general manner without the limits of Wikipedia.

For this purpose, we propose a novel generative model of the semantics of browsing. Referring to the notations in Table 1, we start by positing a distribution over joint latent transition topics (as a higher level semantic space), θ_t for each transition t. The corresponding source, $t(s)$ and destination, $t(d)$ articles of a given transition t are assumed to be admixtures of latent topics which are conditioned on the joint topic transition distribution, θ_t. The detailed generative process is given in the following subsection. For ease of reference, we will refer to this model as the Joint Transition Topic Model (**JTT**).

5.2. Generative Process

We now detail the full generative process of our Joint Transition Topic Model. The variable names and their descriptions are provided in Table 1. Figure 2 shows the plate notation of our model.

Variable	Description
t	A transition t
$t(s), t(d)$	The source and destination pages of t
$\theta_t \sim Dir(\alpha)$	Joint source/destination topic distribution
z_s, z_d	Latent topics of $t(s), t(d)$ respectively
w_s, w_d	Observed word tokens of $t(s), t(d)$ respectively
$\varphi_k \sim Dir(\beta)$	Latent topic-word distributions for topic k
α, β	Dirichlet parameters for θ, φ
N_s, N_d	No. of terms in source and destination pages of t
$T = \{t\}$	Set of all transitions, t
K	No. of topics
V	No. of unique terms in the vocabulary
Z^S, Z^D	Set of all topics in source, destination pages
W^S, W^D	Set of all word token in source, destination pages
$\Theta = \{\theta_t\}$	Set of all latent joint transition topic distributions
$\Phi = \{\varphi_k\}$	Set of all latent topics
$\theta_{t,k}$	Contribution of topic k in transition t
$w^S_{t,j}, w^D_{t,j}$	jth word of transition t in $t(s), t(d)$
$z^S_{t,j}, z^D_{t,j}$	Latent topic of jth word of transition t in $t(s), t(d)$
$n^S_{t\ s\ ,k}$	No. of words in $t(s)$ assigned to topic k
$n^D_{t\ d\ ,k}$	No. of words in $t(d)$ assigned to topic k
$n^S_{k,v}$	No. of times word v assigned to topic k in W^S
$n^D_{k,v}$	No. of times word v assigned to topic k in W^D

Table 1: List of notations

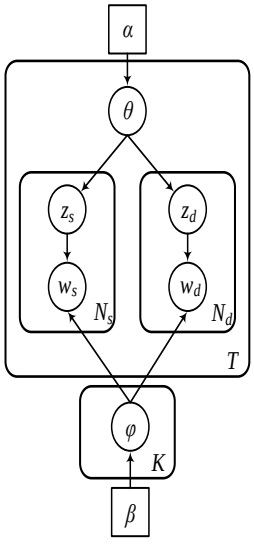

Figure 2: Plate Notation of JTT.

For each topic k, draw $\varphi_k \sim Dir(\beta)$

1) For each transition t:
   a) Draw the joint topic transition distribution, $\theta_t \sim Dir\ \alpha$
   b) For each word token $j \in \{1 \ldots N_S\}$:
      i) Draw $z^S_{t,j} \sim Mult\ \theta_t$
      ii) Emit $w^S_{t,j} \sim Mult\ \varphi_k$
   c) For each word token $j \in \{1 \ldots N_D\}$:
      i) Draw $z^D_{t,j} \sim Mult\ \theta_t$
      ii) Emit $w^D_{t,j} \sim Mult(\varphi_k)$

5.3. Inference

This section details the model inference. We start by deriving the joint distribution of our model and then employ Markov Chain Monte Carlo (MCMC) Gibbs sampling for learning the model from data. To derive the joint distribution, we factor our model according to the causalities governed by the Bayesian network of the JTT model.

$$P\ W^S, W^D, Z^S, Z^D\ = P\ W^S, W^D | Z^S, Z^D\ P(Z^S, Z^D)$$

We employ approximate posterior inference using Monte Carlo Gibbs sampling and use Rao-Blackwellization to reduce sampling variance by collapsing on the latent variables θ and φ. We first compute the second factor as follows:

$$P\ Z^S, Z^D\ = \int P(Z^S, Z^D|\Theta) d\Theta$$

$$= \int \prod_{t=1}^{T} \left\{ \left(\prod_{j=1}^{N_s} p(z^S_{t,j}|\theta_t) \right) \left(\prod_{j=1}^{N_d} p(z^D_{t,j}|\theta_t) \right) \left(\frac{1}{B\ \alpha} \prod_{k=1}^{K} (\theta_{t,k})^{\alpha-1} \right) \right\} d\Theta$$

$$= \int \prod_{t=1}^{T} \left\{ \begin{array}{c} \left(\prod_{j=1}^{N_s} \prod_{k=1}^{K} (\theta_{t,k})^{x^{t,j}_k} \right) \times \\ \left(\prod_{j=1}^{N_d} \prod_{k=1}^{K} (\theta_{t,k})^{x^{t,j}_k} \right) \left(\frac{1}{B(\alpha)} \prod_{k=1}^{K} (\theta_{t,k})^{\alpha-1} \right) \end{array} \right\} d\Theta$$

where $x^{t,j}_k = [x^{t,j}_1, x^{t,j}_2, \ldots\ x^{t,j}_K]$ is 1-of-K encoded and exactly one of its component attains a value of 1 while rest other 0. Upon careful observation and groupings, these can be replaced with appropriate count variables as follows:

$$\int \prod_{t=1}^{T} \left(\frac{1}{B(\alpha)} \prod_{k=1}^{K} (\theta_{t,k})^{n^S_{t\ s\ ,k} + n^D_{t\ d\ ,k} + \alpha - 1} \right) d\Theta$$

As each θ_t is conditionally independent given the hyperparameter α, we have:

$$\prod_{t=1}^{T} \frac{1}{B\ \alpha} \int \left(\prod_{k=1}^{K} (\theta_{t,k})^{n^S_{t\ s\ ,k} + n^D_{t\ d\ ,k} + \alpha - 1} \right) d\theta_t$$

$$= \prod_{t=1}^{T} \frac{B(n^S_{t\ s\ ,[\,]} + n^D_{t\ d\ ,[\,]} + \alpha)}{B(\alpha)}$$

The above simplification can be obtained using techniques for solving Dirichlet integrals (Box & Tiao, 1973) and consequently reduced. Omission of a latter index in the count variables denoted by [ ] corresponds to the row vector spanning over the latter index. The function $B(\cdot)$ refers to the multinomial beta function given below:

$$B\ \alpha\ = B(\alpha_1, \ldots \alpha_{\dim\ \alpha}\ ) = \frac{\prod_{i=1}^{\dim(\alpha)} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{\dim(\alpha)} \alpha_i)}$$

Using similar techniques as above, it is not hard to show that the first factor simplifies as follows:

$$P\ W^S, W^D | Z^S, Z^D\ = \prod_{t=1}^{K} \frac{B(n^S_{k,[\,]} + n^D_{k,[\,]} + \beta)}{B(\beta)}$$

Thus, the full joint can be written succinctly as follows:

$$P\ W^S, W^D, Z^S, Z^D = \left(\prod_{t=1}^{T} \frac{B(n^S_{t\ s\ ,[\,]}\ +\ n^D_{t\ d\ ,[\,]} +\ \alpha)}{B(\alpha)} \right) \left(\prod_{t=1}^{K} \frac{B(n^S_{k,[\,]}\ +\ n^D_{k,[\,]} + \beta)}{B(\beta)} \right)$$

Having obtained the joint, it is straightforward to derive the Gibbs sampler for learning the model. We start with the Gibbs conditional distribution for z^S.

$$\begin{aligned} p(z^S_{t,j} = k | W^S_{\neg\ t,j}, Z^S_{\neg\ t,j}, W^D, Z^D, w^S_{t,j}) &\propto \frac{P\ W^S, W^D, Z^S, Z^D}{p(w^S_{t,j}) P\left(W^S_{\neg\ t,j},\ W^D | Z^S_{\neg\ t,j}, Z^D\right) P\left(Z^S_{\neg\ t,j}, Z^D\right)} \\ &\propto \frac{\left[\left(\prod_{t=1}^{T} \frac{B(n^S_{t\ s\ ,[\,]}\ +\ n^D_{t\ d\ ,[\,]} +\ \alpha)}{B(\alpha)} \right) \left(\prod_{t=1}^{K} \frac{B(n^S_{k,[\,]}\ +\ n^D_{k,[\,]} + \beta)}{B(\beta)} \right) \right]}{\left[\left(\prod_{t=1}^{T} \frac{B(n^S_{t\ s\ ,[\,]}\ +\ n^D_{t\ d\ ,[\,]} +\ \alpha)}{B\ \alpha} \right) \left(\prod_{t=1}^{K} \frac{B(n^S_{k,[\,]}\ +\ n^D_{k,[\,]} + \beta)}{B\ \beta} \right) \right]_{\neg\ t,j}} \\ &\propto \left(\prod_{t=1}^{T} \frac{B(n^S_{t\ s\ ,[\,]}\ +\ n^D_{t\ d\ ,[\,]} +\ \alpha)}{B(n^S_{t\ s\ ,[\,]}\ +\ n^D_{t\ d\ ,[\,]} +\ \alpha)_{\neg\ t,j}} \right) \left(\prod_{k=1}^{K} \frac{B(n^S_{k,[\,]}\ +\ n^D_{k,[\,]} + \beta)}{B(n^S_{k,[\,]}\ +\ n^D_{k,[\,]} + \beta)_{\neg\ t,j}} \right) \end{aligned}$$

where the subscript $(\neg\ t, j\ )$ denotes the value of the expression excluding the counts of the term t, j .

Further, noting that

$$\left(n^S_{t\ s\ ,k}\right)_{\neg\ t,j} = \begin{cases} n^S_{t\ s\ ,k}, & k \neq k' \\ n^S_{t\ s\ ,k} - 1, & k = k' \end{cases}$$

where k' is a candidate topic assigned to t, j , we can simplify the above expression further as follows:

$$p(z^S_{t,j} = k | \ldots) \propto \frac{(n^S_{t\ s\ ,k})_{\neg\ t,j}\ +\ n^D_{t\ d\ ,k} + \alpha}{\sum_{k=1}^{K} \left(\left(n^S_{t\ s\ ,k}\right)_{\neg\ t,j}\ +\ n^D_{t\ d\ ,k} + \alpha \right)} \frac{(n^S_{k,v})_{\neg\ t,j}\ +\ n^D_{k,v} + \beta}{\sum_{v=1}^{V} \left((n^S_{k,v})_{\neg\ t,j}\ +\ n^D_{k,v} + \beta \right)}$$

Using derivation techniques similar to above, one can derive the Gibbs conditional for z^D as follows:

$$p(z^D_{t,j} = k | \ldots) \propto \frac{n^S_{t\ s\ ,k}\ + \left(n^D_{t\ d\ ,k}\right)_{\neg\ t,j} + \alpha}{\sum_{k=1}^{K} \left(n^S_{t\ s\ ,k}\ + \left(n^D_{t\ d\ ,k}\right)_{\neg\ t,j} + \alpha \right)} \frac{n^S_{k,v}\ +\ (n^D_{k,v})_{\neg\ t,j}\ + \beta}{\sum_{v=1}^{V} \left(n^S_{k,v}\ +\ \left(n^D_{k,v}\right)_{\neg\ t,j} + \beta \right)}$$

5.4. Posterior

Having sampled latent variable $\{z^S, z^D\}$ and observed$\{W^S, W^D\}$, we now detail the posterior predictive distribution for our model. Using the fact that Dirichlet is conjugate to the Multinomial, the point estimates of the posterior distributions can be computed as follows:

$$\widehat{\theta_{t,k}} = E[\theta_{t,k} | Z^S, Z^D] = \frac{n^S_{t\ s\ ,k}\ +\ n^D_{t\ d\ ,k} +\ \alpha}{\sum_{k=1}^{K} \left(n^S_{t\ s\ ,k}\ +\ n^D_{t\ d\ ,k} +\ \alpha \right)}$$

$$\widehat{\varphi_{k,v}} = E[\varphi_{k,v}|W^S, W^D] = \frac{n^S_{k,v} + n^D_{k,v} + \beta}{\sum_{v=1}^{V}(n^S_{k,v} + n^D_{k,v} + \beta)}$$

5.5. Training Settings

We learned our joint topic model from a random traffic-weighted sample of 10,000 transition comprising of source and destination pages. The Dirichlet hyperparameters were set to α = 50/κ and β = 0.1 according to the values suggested in (Griffiths, 2004). The number of topics, κ, was empirically set to 50. We also conducted pilot experiments with other hyperparameter settings and larger data-sets and more topics. However, these exploratory experiments with varying settings did not make a substantial difference. Although increasing the number of topics and modeling more volume usually results in lowering of perplexities and better fitting in topic models (Blei et al., 2003), it can also result in redundancy in topics which may not be very useful for downstream applications (Chen et al., 2013). In our setting, where the eventual downstream task is modeling interestingness, our pilot experiments showed that the value of information contained in the latent semantics of our model reached saturation upon increasing number of topics and the data size. For all reported experiments we used the posterior estimates of our joint model learned according to the above mentioned settings.

In our interestingness prediction model we use three classes of features from the joint topic model, capturing the topic distribution of source, destination, and transition: source topic features (Z^S, labeled as **JTT$_{src}$** in charts), destination topic features (Z^D, labeled as **JTT$_{dst}$**), and transition topic features (Θ, labeled as **JTT$_{trans}$**). The value of each topic feature is the probability of the topic.

6. Experiments

We evaluate our models on two tasks: *highlighting* and *click prediction.*

6.1. Results: Highlighting

***Highlights*. Propose *k* anchors that the user will find interesting**. In this task, a user is reading a document $s \in D$ and is interested in learning more about a set of anchors. We formally define the interest function:

$$\boldsymbol{\varphi}: \boldsymbol{D} \times \boldsymbol{A} \rightarrow \mathbb{R} \tag{2}$$

where $\varphi\ s, a$ reflects the user's degree of interest in *a* while consuming $s \in D$.

Our goal in this task is to select *k* anchors that maximize the cumulative degree of interest of the user. We consider the ideal selection to consists of the *k* most interesting anchors, ranked in decreasing order of $\varphi\ s, a$.

We instantiate our interest models on this task by scoring each source-anchor pair and by ordering the anchors in decreasing order of their prediction score. We then select the top-*k* for each source document. Given a source document *s*, we measure the quality of a system's ranking against the ideal ranking defined above using the standard nDCG metric 0. In order to compute nDCG, we need to assign a relevance score to each anchor a proposed by our model. We consider the degree of interest of the user in a as the best measure of relevance, hence we use $\varphi\ s, a$.

Table 2: nDCG results for different feature sets across HEAD, TORSO, and TAIL. Bold indicates statistically significant best systems (with 95% confidence).

nDCG %	HEAD				TORSO				TAIL			
	@1	@2	@5	@10	@1	@2	@5	@10	@1	@2	@5	@10
Baseline: random	4.07	4.90	6.24	8.10	3.56	4.83	7.66	10.92	6.20	11.74	19.50	25.82
Baseline: first n anchors	9.99	12.47	17.72	24.33	7.17	9.87	17.06	23.97	9.06	16.66	27.35	34.82
Anc	21.46	22.50	25.30	29.47	13.85	16.80	22.85	28.20	10.88	19.16	29.33	36.48
Anc+GeoTemp	18.99	21.26	25.06	29.80	13.70	16.54	22.52	27.98	10.72	18.90	29.19	36.35
Anc+JTT_{dst}	13.97	16.33	19.69	23.78	11.37	14.17	19.67	24.66	11.62	19.69	29.69	36.35
Anc+JTT_{dst}+JTT_{src}	26.62	30.03	34.82	39.38	17.05	20.82	**27.15**	**32.48**	12.27	**21.56**	**31.88**	**38.85**
Anc+JTT_{dst}+JTT_{src}+JTT_{trans}	**34.49**	**35.21**	**38.01**	**41.80**	**18.32**	**21.69**	**28.03**	**33.22**	**13.06**	**21.68**	**32.13**	**39.01**

Computing $\varphi\ s,a\ $ is of course a challenge. Recall our test set described in Section 4.1. Each item in the test set consists of a real-world browsing transition denoted $\tau\ s,a,t\ $, where a represents the anchor clicked from document s that led to document t. In order to approximate $\varphi\ s,a\ $, we compute the aggregate average clicks on a from the source page s, that is:

$$\varphi\ s,a\ = \frac{|\tau\ s,a,t\ |}{|\tau\ s,*,*\ |}$$

where $|\tau\ s,a,t\ |$ is the number of transitions from s to t via a in the test set and $|\tau\ s,*,*\ |$ is the number of transitions in the test set originating from document s.

Table 2 shows the nDCG results for two baselines and a range of different feature sets. The first high-level observation is that the difficulty of the interestingness prediction problem becomes obvious from the two baseline results. Since there are many anchors on an average page, picking a random set of anchors yields very low nDCG scores. Note also that nDCG numbers of our baselines increase as we move from HEAD to TORSO to TAIL, due to the fact that the average number of links per page decreases in these sets from 170 to 94 to 41. The second baseline illustrates that it is not by any means sufficient to simply pick the top *n* anchors on a page. Using our set of anchor features as described in Section 4.2 in a regression model greatly improves over the baselines, with the strongest numbers on the HEAD set and decreasing effectiveness in TORSO and TAIL. This shows that the distribution of interesting anchors on a page actually differs according to the popularity of the content, possibly also with the length of the page since popular pages also tend to attract more time from editors and often have longer content. Adding geotemporal features to the anchor features leads to slightly degraded results. This is despite the fact that geotemporal features in our initial exploration in Section 3.1.2 showed some impact. We believe that this might be due to overtraining, possibly because of the high number of observed values for categorical features such as user city etc. Not shown in Table 2 are additional experiments that consistently showed no positive impact from the geotemporal features in combination with other feature sets. Based on these results, we excluded this feature set from subsequent experiments. Our best performing model is the one using anchor features and all three sets of latent semantic features (source, destination, and transition topics). The biggest improvement is obtained on the HEAD

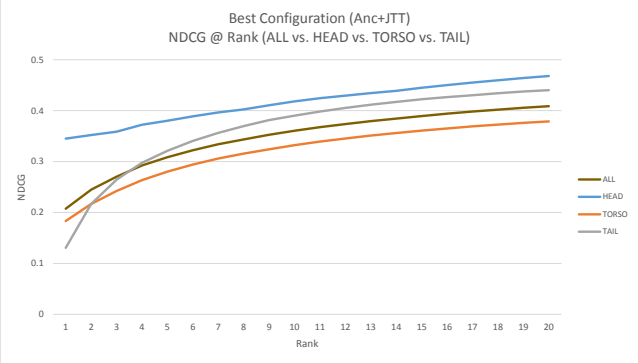

Figure 3: Overall performance (ALL) versus HEAD, TORSO, and TAIL subsets.

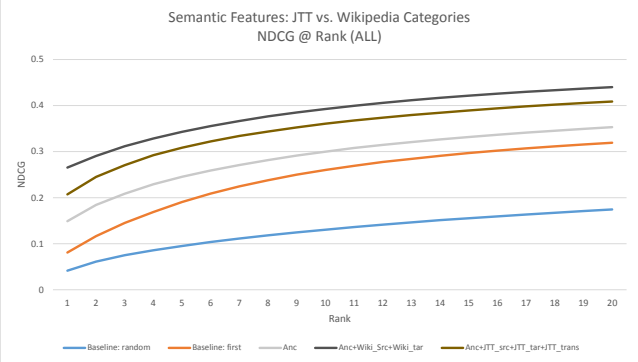

Figure 4: JTT features versus Wikipedia category features over complete test set.

data. This is not surprising given that the topic model is trained on a traffic weighted sample of Wikipedia articles and that HEAD pages tend to have more content, making the identification of topics more reliable. Regarding the individual contributions of the latent semantic destination, source and transition features, the observation is that destination features alone hurt performance on the HEAD set. Latent semantic source features lead to a boost across the board, and the addition of latent semantic transition topic features produces the best model, with gains especially pronounced on the HEAD data. Figure 3 shows the performance of our best configuration across ALL, HEAD, TORSO, and TAIL. Interestingly, the TAIL exhibits better performance of the model than the TORSO (with the exception of nDCG at rank 3 or higher). We attribute this to the observation that the average number of anchors in a TAIL page is less than half of that in a TORSO page.

6.1.1. The Contribution of Semantic Features

In further experiments we addressed the question how our unsupervised latent semantic features perform compared to editor assigned topics for Wikipedia pages. This is an important question for two reasons. First, it is reasonable to consider the manually assigned topics as a (fine-grained) oracle for topic assignments and hence as a good comparison for any topic model built on the same data. Each article is assigned multiple categories at the discretion of the editor. New categories are constantly added

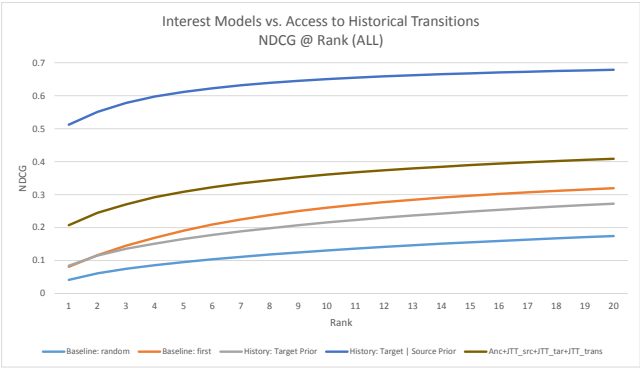

Figure 5: Comparison of best configuration interest model against baselines and access to historical transitions.

to Wikipedia, and the total number is in the hundreds of thousands. Second, note that outside of Wikipedia, the luxury of manually assigned categories/topics does not exist, so it is important to see for a model of interestingness how much of the information in manual categories can be recovered through an unsupervised topic model. Figure 4 illustrates that Wikipedia categories outperform the JTT topic features, but the latter can recover about two thirds of the nDCG gain from Wikipedia categories.

6.1.2. Access to History

For the HEAD part of the data, we have enough historical clickthrough data that we can leverage directly for prediction. To establish how well this strategy fares we conducted experiments where we used the prior probability $p(d|s)$ obtained from the training data (both smoothed and unsmoothed). Following this strategy we can achieve up to 65% nDCG@10 as shown in Figure 5 where the use of prior history is compared to our best model and to baselines. In a real-life application scenario, this is not a viable option, though. First, it is not suitable for a "cold start" scenario where we need to determine interestingness for a new, unseen page. More importantly, the TORSO and TAIL data sets have no or only very sparse histories, hence we cannot resort to this information. Finally, clickthrough history makes sense in our experimental setting where Wikipedia anchors and destination pages are static. Any application which identifies interesting nuggets in order to trigger a query for that nugget on a search engine would not be able to rely on a static result, but would retrieve dynamically produced and ranked content which may, for example, contain a link to the latest gossip news which did not exist only a few hours before.

6.2. Results: Click Prediction

Our *highlights* task reflects the main goal of our paper, i.e., to predict interestingness in the context of any document, whether it be a web page, an email, or a book. A natural extension of our work, especially in our experimental setting with Wikipedia transitions, is to predict the next click of a user, i.e., click prediction.

There is a subtle but important difference between the two tasks. *Highlights* aims to identify a set of interesting nuggets for a source document. A user may ultimately click on only a subset of the nuggets, and perhaps not in the order of most interest. Our experimental metric, nDCG, reflects this ranking task well. Click prediction is an inherently more difficult task, where we focus on predicting exactly the next

Table 3: Click prediction results for different feature sets across HEAD, TORSO, and TAIL. Bold indicates statistically significant best systems (with 95% confidence).

Cumulative Precision %	HEAD				TORSO				TAIL			
	@1	@2	@5	@10	@1	@2	@5	@10	@1	@2	@5	@10
Baseline: random	1.07	2.08	5.29	10.55	1.94	3.91	9.71	19.00	5.97	11.66	26.43	44.94
Baseline: first *n* anchors	2.68	5.77	16.73	33.78	4.10	8.19	22.86	42.08	8.77	16.57	36.80	58.52
Anc	8.40	12.55	22.04	34.22	8.70	14.37	27.56	42.68	10.59	19.08	38.27	59.04
Anc+GeoTemp	7.14	11.77	21.81	35.21	8.39	14.00	27.12	42.58	10.42	18.80	38.25	59.11
Anc+JTT_{dst}	5.48	9.19	17.77	29.14	6.93	12.07	23.90	38.00	11.23	19.59	38.46	57.87
Anc+JTT_{dst}+JTT_{src}	9.02	15.65	30.05	**44.72**	10.11	17.42	32.08	**47.07**	**11.95**	**21.47**	**40.96**	**61.24**
Anc+JTT_{dst}+JTT_{src}+JTT_{trans}	**11.53**	**18.43**	**31.93**	**45.36**	**10.86**	**18.19**	**32.96**	**47.66**	**12.64**	**21.58**	**41.27**	**61.28**

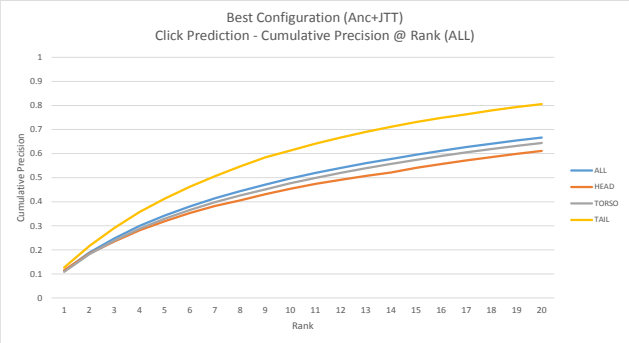

Figure 6: Overall performance (ALL) versus HEAD, TORSO, and TAIL subsets.

click of a specific user. Unlike in the highlights task, there is no partial credit for retrieving other interesting anchors. Only the exact clicked anchor is considered a correct result. As such, we utilize a different metric than nDCG on this task. We measure our model's performance on the task of click prediction using cumulative precision. Given a unique transition event $\tau(s, a, d)$ by a particular user at a particular time, we present the transition, minus the gold anchor a and destination d, to our models, which in turn predict an ordered list of most likely anchors on which the user will click. The cumulative precision at k of a model, is 1 if any of the predicted anchors matched a, and 0 otherwise.

Table 3 outlines the results on this task and Figure 6 shows the corresponding chart for our best configuration. Note that in the click prediction task, the model performs best on the TAIL, followed by TORSO and HEAD. This seems to be a reflection of the fact that in this harder task, the total number of anchors per page is the most influential factor in model performance.

7. Conclusion and Future Directions

We presented a notion of interestingness of a text nugget on a page that is grounded in observable behavior during content consumption. We implemented a model for prediction of interestingness of anchor texts that we trained and tested within the domain of Wikipedia. The model design is not tied to our experimental choice of using Wikipedia and can be applied to other domains. Our model takes

advantage of semantic features that we derive from a powerful and novel joint topic model. The joint topic model takes into account topic distributions for source, destination and transitions from source to destination. We demonstrated that the latent semantic features from our topic model contribute significantly to the performance of interestingness prediction, to the point where they perform nearly as well as using manually assigned Wikipedia categories as features. We also showed that the transition topics improve results over just using source and destination semantic features alone.

A number of future directions immediately suggest themselves given the discussion and results in this paper. First of all, we did not address the problem of generating a candidate list of nuggets or text strings. In our problem setting, the candidate list is the list of anchor texts on a page. For an application that marks interesting nuggets on an arbitrary page, however, we would need a detector for nugget candidates. A simple first approach would be to use a state-of-the-art Named Entity Recognition (NER) system. This does not solve the problem entirely, since we know that named entities are not the only interesting nuggets – general terms and concepts can also be of interest to a reader. On the other hand we do have reason to believe that entities play a very prominent role in web content consumption, based on the frequency with which entities are searched for (see, for example 0 and the references cited therein). Recall also from Section 3.1.1 that we made similar observations in the Wikipedia domain where "Creative Work", "Organization", "Person" and "Place" pages attracted most of the traffic. Using an NER system as a candidate generator would also allow us to add another potentially useful feature to our interestingness prediction model: the type of the entity. One could also envision a model of interestingness and nugget detection that tackles both problems as a joint objective instead of a two-stage pipelined process. A second point concerns the observation from the previous section on the different regularities that seem to be at play according to the popularity and possibly the length of an article. More detailed experiments are needed to tease out this influence and possibly improve the predictive power of the model. It would also be interesting to revisit the user/time features which in our results led to overfitting. Pruning categorical features to a smaller number of observed values or eliminating the particularly fine-grained features such as user city could possibly still benefit the model. Finally, there are a number of options regarding our joint topic model that could be explored further. Being trained on a traffic-weighted sample of articles, the topic model predominantly picks up on popular topics. This could be remedied by training on a non-weighted sample, or, more promisingly, on a larger non-weighted sample with a larger K, i.e. more permissible total topics. We are also considering training different topic models for the HEAD, TORSO and TAIL portions of our data to determine if we can improve especially TORSO and TAIL performance with these "customized" topic models.

References

Agichtein, E., Brill, E., and Dumais, S. 2006. Improving web search ranking by incorporating user behavior information. *Proceedings of SIGIR*, 19-26.

Bandari, R., Asur, S., and Huberman, B. A. 2012. The Pulse of News in Social Media: Forecasting Popularity. *Proceedings of ICWSM*.

Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., and Su, Z. 2007. Optimizing web search using social annotations. *Proceedings of WWW*, 501-510.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993-1022.

Box, G. E., and Tiao, G.C. 1973. Bayesian inference in statistical analysis. Addison-Wesely.

Broder, A., Fontoura, M., Josifovski, V., and Riedel, L. 2007. A semantic approach to contextual advertising. *Proceedings of SIGIR*, 559-566.

Buscher, G., van Elst, L., and Dengel, A. 2009. Segment-level display time as implicit feedback: a comparison to eye tracking. *Proceedings of SIGIR*, 1-30.

Chang, J., Boyd-Graber, J., and Blei, D. M. 2009. Connections between the lines: augmenting social networks with text. *Proceedings of KDD*, 169-178.

Chakrabarti, D., Agarwal, D., and Josifovski, V. 2008. Contextual advertising by combining relevance with click feedback. *Proceedings of WWW*, 417-426.

Chatterjee, P., Hoffman, D. L., and Novak, T. P. 2003. Modeling the clickstream: Implications for web-based advertising efforts. *Marketing Science*, *22*(4), 520-541.

Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M. and Ghosh, R. 2013. Leveraging Multi-Domain Prior Knowledge in Topic Models. *Proceedings of IJCAI*, 2071-2077.

Chen, Y., and Yan, T. W. 2012. Position-normalized click prediction in search advertising. *Proceedings of KDD*, 795-803.

Cheng, H. and Cantú -Paz, E. 2010. Personalized click prediction in sponsored search. *Proceedings of WSDM*, 351-360.

Cheng, H., Zwol, R. V., Azimi, J., Manavoglu, E., Zhang, R., Zhou, Y., and Navalpakkam, V. 2012. Multimedia features for click prediction of new ads in display advertising. *Proceedings of KDD*, 777-785.

Craswell, N., Zoeter, O., Taylor, M., and Ramsey, B. 2008. An experimental comparison of click position-bias models. *Proceedings of WSDM*, 87-94.

Dhillon, I. S., and Modha, D. S. 2001. Concept decompositions for large sparse text data using clustering. *Machine learning*, 42(1-2), 143-175.

Erosheva, E., Fienberg, S., and Lafferty, J. 2004. Mixed membership models of scientific publications. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1), 5220-5227.

Evans, B. and Chi, E. 2010. An Elaborated Model of Social Search. *Information Processing and Management,* 46(6):656-678.

Farahat, A., and Bailey, M. C. 2012. How effective is targeted advertising? *Proceedings of WWW*, 111-120.

Friedman, J. H. 1999. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189-1232, 1999.

Gao, J., Toutanova, K., and Yih, W. T. 2011. Clickthrough-based latent semantic models for web search. *Proceedings of SIGIR*, 675-684.

Ghosh, A. and Mahdian, M. 2008. Externalities in online advertising. *Proceedings of WWW*, 161-168.

Graepel, T., Candela, J.Q., Borchert, T., and Herbrich, R. 2010. Web-scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine. *Proceedings of ICML*, 13-20.

Granka, L. A., Joachims, T., and Gay, G. 2004. Eye-tracking analysis of user behavior in www search. *Proceedings of SIGIR*, 478-479.

Griffiths, T.L. 2004. Finding Scientific Topics. *Proceedings of the National Academy of Science*, 101, suppl 1, 5228-5235.

Guo, Q., and Agichtein, E. 2012. Beyond Dwell Time: Estimating Document Relevance from Cursor Movements and other Post-click Searcher Behavior. *Proceedings of WWW*, 569-578.

Guo, F., Liu, C., Kannan, A., Minka, T., Taylor, M., Wang, Y.-M, and Faloutsos, C. 2009. Click chain model in web search. *Proceedings of WWW*, 11-20.

Guo, J., Xu, G., Cheng, X., and Li, H. 2009. Named entity recognition in query. *Proceedings of SIGIR*, 267-274.

Hillard, D., Manavoglu, E., Raghavan, H., Leggetter, C., Cantú-Paz, E., and Iyer, R. 2011. The sum of its parts: reducing sparsity in click estimation with query segments. *Information Retrieval Journal*, 14(3), 315-336. Springer, Berlin Heidelberg.

Huang, A., Milne, D., Frank, E., and Witten, I. H. 2009. Clustering Documents Using a Wikipedia-Based Concept Representation. *Advances in Knowledge Discovery and Data Mining*, 628-636. Springer, Berlin Heidelberg.

Joachims, T. 2002. Optimizing search engines using clickthrough data. *Proceedings of KDD*, 133-142.

Joachims, T., Granka, L., Pan, B., Hembrooke, H. and Gay, G. 2005. Accurately interpreting clickthrough data as implicit feedback. *Proceedings of SIGIR*, 154-161.

Kelly, D. and Belkin, N. J. 2004. Display time as implicit feedback: understanding task effects. *Proceedings of SIGIR*, 377-384.

Lacerda, A., Cristo, M., Gonçalves, M. A., Fan, W., Ziviani, N., and Ribeiro-Neto, B. 2006. Learning to advertise. *Proceedings of SIGIR*, 549-556.

Lerman, K., and Hogg, T. 2010. Using a model of social dynamics to predict popularity of news. *Proceedings of WWW*, 621-630.

Lin, T., Pantel, P., Gamon, M., Kannan, A., and Fuxman, A. 2012. Active objects: actions for entity-centric search. *Proceedings of WWW*, 589-598.

Maedche, A., and Staab, S. 2000. Mining ontologies from text. *Knowledge Engineering and Knowledge Management Methods, Models, and Tools*, 189-202. Springer, Berlin Heidelberg.

Manning, C. D., Raghavan, P., and Schutze, H. 2008. Introduction to Information Retrieval. Cambridge University Press.

McKeown, K. R., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J. L., Nenkova, A., Sable, C., Schiffman, B., and Sigelman, S. 2002. Tracking and summarizing news on a daily basis with Columbia's Newsblaster. *Proceedings of HLT*, 280-285.

McMahan, H. B., Holt, G., Sculley, D., Young, M., Ebner, D., Grady, J., and Kubica, J. 2013. Ad click prediction: a view from the trenches. *Proceedings of KDD*, 1222-1230.

Metzler, D., and Croft, W. B. 2007. Latent concept expansion using markov random fields. *Proceedings of SIGIR*, 311-318.

Morita, M. and Shinoda, Y. 1994. Information filtering based on user behavior analysis and best match text retrieval. *Proceedings of SIGIR*, 272-281.

Okumura, A., and Hovy, E. 1994. Lexicon-to-ontology concept association using a bilingual dictionary. *Proceedings of the First Conference of the Association for Machine Translation in the Americans*, 177-184.

Ramage, D., Heymann, P., Manning, C. D., and Garcia-Molina, H. 2009. Clustering the tagged web. *Proceedings of WSDM*, 54-63.

Richardson, M., Dominowska, E., and Ragno, R. 2007. Predicting clicks: estimating the click-through rate for new ads. *Proceedings of WWW*, 521-530.

Sanderson, M., and Croft, B. 1999. Deriving concept hierarchies from text. *Proceedings of SIGIR*, 206-213.

Shen, S., Hu, B., Chen, W., and Yang, Q. 2012. Personalized click model through collaborative filtering. *Proceedings of WSDM, 323-333.*

Szabo, G., and Huberman, B. A. 2010. Predicting the popularity of online content. *Communications of the ACM*, *53*(8), 80-88.

Wang, X., Broder, A., Fontoura, M., and Josifovski, V. 2009. A search-based method for forecasting ad impression in contextual advertising. Proceedings of *WWW*, 491-500.

Wang, H., Dong, A., Li, L., Chang, Y., and Gabrilovich, E. 2012. Joint relevance and freshness learning from clickthroughs for news search. *Proceedings of WWW*, 579-588.

White, R. W. and Kelly, D. 2006. A study on the effects of personalization and task information on implicit feedback performance. *Proceedihngs of CIKM*, 297-306.

Xiong, C., Wang, T., Ding, W., Shen, Y., and Liu, T. Y. 2012. Relational click prediction for sponsored search. *Proceedings of WSDM*, 493-502.

Xu, D., Liu, Y., Zhang, M., Ma, S., and Ru, L. 2012. Incorporating revisiting behaviors into click models. *Proceedings of WSDM*, 303-312.

Xu, H., Gao, B., Yang, D., and Liu, T. Y. 2013. Predicting advertiser bidding behaviors in sponsored search by rationality modeling. *Proceedings of WWW*, 1433-1444.

Xu, W., Manavoglu, E., and Cantú-Paz, E. 2010. Temporal click model for sponsored search. *Proceedings of SIGIR*, 106-113.

Yano, T., Cohen, W. W., & Smith, N. A. 2009. Predicting response to political blog posts with topic models. *Proceedings of NAACL*, 477-485.