Predicting Interesting Things in Text

Michael Gamon
Microsoft Corp.
One Microsoft Way
Redmond, WA 98052
mgamon@microsoft.com

Arjun Mukherjee
Department of Computer Science
University of Houston
Houston, TX 77004
arjun4787@gmail.com

Patrick Pantel
Microsoft Corp.
One Microsoft Way
Redmond, WA 98052
ppantel@microsoft.com

Abstract

While reading a document, a user may encounter concepts, entities, and topics that she is interested in exploring more. We propose models of "interestingness", which aim to predict the level of interest a user has in the various text spans in a document. We obtain naturally occurring interest signals by observing user browsing behavior in clicks from one page to another. We cast the problem of predicting interestingness as a discriminative learning problem over this data. We leverage features from two principal sources: textual context features and topic features that assess the semantics of the document transition. We learn our topic features without supervision via probabilistic inference over a graphical model that captures the latent joint topic space of the documents in the transition. We train and test our models on millions of real-world transitions between Wikipedia documents as observed from web browser session logs. On the task of predicting which spans are of most interest to users, we show significant improvement over various baselines and highlight the value of our latent semantic model.

1 Introduction

Reading inevitably leads people to discover interesting concepts, entities, and topics. Predicting what interests a user while reading a document has important applications ranging from augmenting the document with supplementary information, to ad placement, to content recommendation. We define the task of predicting **interesting things (ITs)** as ranking text spans in an unstructured document according to whether a user would want to know more about them. This desire to learn more serves as our proxy for interestingness.

There are many types of observable behavior that indicate user interest in a text span. The closest one to our problem definition is found in web browsing, where users click from one document to another via named anchors. The click process is generally governed by the user's interest (modulo erroneous clicks). As such, the anchor name can be viewed as a text span of interest for that user. Furthermore, the frequency with which users, in aggregate, click on an anchor serves as a good proxy for the level of interest[1].

What is perceived as *interesting* is influenced by many factors. The semantics of the document and candidate IT are important. For example, we find that when users read an article about a movie, they are more likely to browse to an article about an actor or character than to another movie or the director. Also, user profile and geo-temporal information are relevant. For example, interests can differ depending on the cultural and socio-economic background of a user as well as the time of the session (e.g., weekday versus weekend, daytime versus late night, etc.).

[1] Other naturally occurring expressions of user interest, albeit less fitting to our problem, are found in web search queries, social media engagement, and others.

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1477–1488, Dublin, Ireland, August 23-29 2014.

Strictly speaking, human interestingness is a psychological and cognitive process (Varela et al., 1991). Clicks and long dwell times are salient observed behavioral signals of interestingness that have been well accepted in the information retrieval literature (Claypool et al., 2001; Mueller and Lockerd, 2001). In this paper, we utilize the observed user's browsing behavior as a supervision signal for modeling interestingness. Specifically, we cast the prediction of ITs as a discriminative learning task. We use a regression model to predict the likelihood of an anchor in a Wikipedia article to be clicked, which as we have seen above can serve as a proxy for interestingness. Based on an empirical study of a sample of our data, we use features in our model from the document context (such as the position of the anchor text, frequency of the anchor text in the current paragraph, etc.) as well as semantic features that aim to capture the latent topic space of the documents in the browsing transition. These semantic features are obtained in an unsupervised manner via a joint topic model of source and target documents in browsing transitions. We show empirical evidence that our discriminative model is effective in predicting ITs and we demonstrate that the automatically learned latent semantic features contribute significantly to the model performance. The main contributions of this paper are:

- We introduce the task of predicting interesting things as identifying what a user likely wants to learn more about while reading a document.
- We use browsing transitions as a proxy for interestingness and model our task using a discriminative training approach.
- We propose a semantic probabilistic model of interestingness, which captures the latent aspects that drive a user to be interested in browsing from one document to another. Features derived from this semantic model are used in our discriminative learner.
- We show empirical evidence of the effectiveness of our model on an application scenario.

2 Related Work

Salience: A notion that might at first glance be confused with interestingness is that of salience (Paranjpe 2009; Gamon et al. 2013). Salience can be described as the centrality of a term to the content of a document. Put another way, it represents what the document is about. Though salience and interestingness can interact, There are clear differences. For example, in a news article about President Obama's visit to Seattle, Obama is salient, yet the average user would probably not be interested in learning more about Obama while reading that article.

Click Prediction: Click prediction models are used pervasively by search engines. Query based click prediction aims at computing the probability that a given document in a search-result page is clicked on after a user enters some query (Joachims, 2002; Joachims et al., 2005; Agichtein et al., 2006; Guo et al., 2009a). Click prediction for online advertising is a core signal for estimating the relevance of an ad to a search result page or a document (Chatterjee et al., 2003; Broder et al., 2007; Craswell et al., 2008; Graepel et al., 2010). Also related are personalized click models, e.g., (Shen et al., 2012), which use user-specific click through rate (CTR). Although these applications and our task share the use of CTR as a supervision signal, there is a key difference: Whereas in web search CTR is used as a predictor/feature at runtime, our task specifically aims at predicting interestingness in the absence of web usage features: Our input is completely unstructured and there is no assumption of prior user interaction data.

Use of probabilistic models: Our semantic model is built over LDA (Blei et al., 2003) and has resemblances to Link-LDA models (Erosheva et al., 2004) and Comment-LDA models (Yano et al., 2009). However, these are tailored for blogs and associated comment discussions which is very different from our source to destination browsing transition logs. Guo et al., (2009b) used probabilistic models for discovering entity classes from query logs and in (Lin et al., 2012), latent intents in entity centric search were explored. Gao et al. (2011) employ statistical machine translation to connect two types of content, learning semantic translation of queries to document titles. None of the above models, however, are directly applicable to the joint topic mappings that are involved in source to destination browsing transitions which are the focus of our work.

Predicting Popular Content: Modeling interestingness is also related to predicting popular content in the Web and content recommenders (Lerman and Hogg, 2010; Szabo and Huberman, 2010; Bandari et al., 2012). In contrast to these tasks, we strive to predict what term a user is likely to be interested in when reading content. We do not rely on prior browsing history, since we aim to predict interestingness in unstructured text with no interaction history. We show in our experiments that a popularity signal alone is not a sufficient predictor for interestingness.

3 The Interestingness Task

The process of identifying interesting things (ITs) on a page consists of two parts: (1) generating candidate things (e.g., entities, concepts, topics); and (2) scoring and ranking these according to interestingness. In this paper, we fix step 1 and focus our effort on step 2, i.e., the assignment of an interestingness score to a candidate. We believe that this scope is appropriate in order to understand the factors that enter into what is perceived as interesting by a user. Once we have gained an understanding of the interestingness scoring problem, however, there are opportunities in identifying candidates automatically, which we leave for future work.

In this section we begin by formally defining our task. We then introduce our data set of naturally occurring interest signals, followed by an investigation of the factors that influence them.

3.1 Formal Task Definition

We define our task as follows. Let U be the set of all documents and A be the set of all candidate text spans in all documents in U, generated by some candidate generator. Let $A_u \subset A$ be the set of candidates in $u \in U$. We formally define the interestingness task as learning the function below, where $\sigma(u, a)$ is the interestingness of candidate a in [2]:

$$\sigma: U \times A \rightarrow \mathbb{R} \quad (1)$$

3.2 Data Set

User browsing events on the web (i.e., a user clicking from one document to another) form a naturally occurring collection of interestingness signals. That is when a user clicks on an anchor in a document, we can postulate that the user is interested in learning more about it, modulo erroneous clicks.

We collect a large set of many millions of such user browsing events from session logs of a commercial web browser. Specifically, we collect from these logs each occurrence of a user click from one Wikipedia page to another during a one month period, from all users in all parts of the world. We refer to each such event as a *transition*. For each transition, our browser log provides metadata, including user profile information, geo-location information and session information (e.g., time of click, source/target dwell time, etc.) Our data set includes millions of transitions between Wikipedia pages.

For our task we require: (1) a mechanism for generating candidate things; (2) ample clicks to serve as a reliable signal of interestingness for training our models; and (3) accessible content. Our focus on Wikipedia satisfies all. First, Wikipedia pages tend to contain many anchors, which can serve as the set of candidate things to be ranked. Second, Wikipedia attracts enough traffic to obtain robust browsing transition data. Finally, Wikipedia provides full content[3] dumps. It is important here to note that our choice of Wikipedia as a test bed for our experiments does not restrict the general applicability of our approach: We propose a semantic model (Section 4.2) for mining latent features relevant to the phenomenon of interestingness which is general and can be applied to generic Web document collections.

Using uniform sampling, we split our data into three sets: a development set (20%), a training set (60%) and a test set (20%). We further subdivide the test set by assigning each transition as belonging to the HEAD, TORSO, or TAIL, which we compute using inverse CDF sampling on the test set. We do so by assigning the most frequently occurring transitions, accounting for 20% of the (source) traffic, to the HEAD. Similarly, the least frequently occurring transitions, accounting for 20% of the (source) traffic, are assigned to the TAIL. The remaining transitions are assigned to the TORSO. This three-way split reflects a common practice in the IR community and is based on the observation that web traffic frequencies show a very skewed distribution, with a small set of web pages attracting a large amount of traffic, and a very long tail of infrequently visited sites. Different regions in that distribution often show marked differences in behaviour, and models that are useful in one region are not necessarily as useful in another.

[2] We fix $\sigma(u, a) = 0$ for all $a \notin A_u$.

[3] We utilize the May 3, 2013 English Wikipedia dump from http://dumps.wikimedia.org, consisting of roughly 4.1 million articles.

3.3 What Factors Influence Interestingness?

We manually inspected 200 random transitions from our development set. Below, we summarize our observations.

Only few things on a page are interesting: The average number of anchors on a Wikipedia page is 79. Of these, only very few are actually clicked by users. For example, the Wikipedia article on the TV series "The Big Bang Theory" leads to clicks on anchors linking to the pages of the series' actors for 90% of transitions (while these anchors account for only a small fraction of all unique anchors on that page).

The semantics of source and destination pages is important: 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 in 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 observed that transitions are influenced by these categories. For example, when the source article category is `Movie`, the most frequently clicked pages are of category `Actor` (63%) and `Character` (13%). For source articles of the `TVSeries` category, `Actor` destination articles account for 86% of clicks. `Actor` articles lead to clicks on `Movie` articles (45%) and other `Actor` articles (26%), whereas `Artist` articles lead to clicks on other `Artist` articles (29%), `Movie` articles (17%) and `MusicRecording` articles (18%).

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 a more prominent position tend to attract more clicks. We noticed similar trends in our data.

The user plays a role: We hypothesized that users from different geographic and cultural backgrounds might exhibit different interests, or that interests are time-bound (e.g., interests on weekends differ from those on week days, daytime from nighttime, etc.) Initial experiments showed small effects of these factors, however, a more thorough analysis on a larger sample is necessary, which we leave for future work.

4 Modeling Interestingness

We cast the problem of learning the interestingness function σ (see Eq. 1) as a discriminative regression learning problem. Below, we first describe this model, and then we introduce our semantic topic model which serves to provide semantic features for the discriminative learner.

4.1 Discriminative Model

Although our task is to predict ITs from unstructured documents, we can leverage the user interactions in our data, described in Section 3.2 as our training signal.

Given a source document $s \in U$, and an anchor in s leading to destination document *d*, we use the aggregate click frequency of this anchor as a proxy for its interestingness, i.e.:

$$\sigma(s,d) = p(d|s) \qquad (2)$$

where $p(d|s)$ is the probability of a user clicking on the anchor to d when viewing s[4]. We use $p(d|s)$ as our regression target computed from our training data.

For our learning algorithm, we use boosted decision trees (Friedman, 1999). We tune our hyperparameters (i.e., number of iterations, learning rate, minimum instances in leaf nodes, and the maximum number of leaves) using cross-validation on the development set. Each transition in our training data is represented as a vector of features, where the features fall into three basic families:

1. Anchor features (**Anc**): position of the anchor in the document, frequency of the anchor, anchor density in the paragraph, and whether the anchor text matches the title of the destination page.
2. User session features (**Ses**): city, country, postal code, region, state and timezone of the user, as well as day of week, hour, and weekend vs. workday of the occurrence of the transition.

[4] For notational convenience, we use $\sigma(s,d)$ even though Eq. 1 defines its second argument as being a candidate text span. Here, it is implicit that *d* consists of both the target document and the anchor text (which serves as the candidate text span).

3 Semantic features: sourced in various experimental configurations from (1) Wikipedia page categories as assigned by Wikipedia editors (**Wiki**) or from (2) an unsupervised joint topic transition model (**JTT**) of source and destination pages (described in detail in the next section).

In some experimental configurations we use Wikipedia page categories as semantic features. We show in our experiments (see Section 5) that these are highly discriminative. It is important to note that editor-labeled category information is available in the Wikipedia domain but not in others. In other words, we can use this information to verify that semantics indeed is influential for interestingness, but we should design our models to not rely on this. We thus build an unsupervised semantic model of source and destination pages, which serves the purpose of providing semantic information without any domain-specific annotation.

4.2 The Semantics of Interestingness

As indicated in Section 3, the semantics of source and destination pages, s and d, influence the likelihood that a user is interested in d after viewing s. In this section we propose an unsupervised method for modeling the transition semantics between s and d. As outlined in the previous section, this model then serves to generate semantic features for our discriminative model of interestingness.

Referring to the notations in Table 1, we start by positing a distribution over the joint latent transition topics (in the higher level of 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 that are conditioned on the joint topic transition distribution, θ_t. For ease of reference, we will refer to this model as the Joint Transition Topic Model (**JTT**). The variable names and their descriptions are provided in Table 1. Figure 1 shows the plate notation of our model and the generative process:

1. For each topic k, draw $\varphi_k \sim Dir(\beta)$
2. 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 $z_{t,j}^S \sim Mult(\theta_t)$
      ii. Emit $w_{t,j}^S \sim Mult(\varphi_k)$
   c. For each word token $j \in \{1 \ldots N_D\}$:
      i. Draw $z_{t,j}^D \sim Mult(\theta_t)$
      ii. Emit $w_{t,j}^D \sim Mult(\varphi_k)$

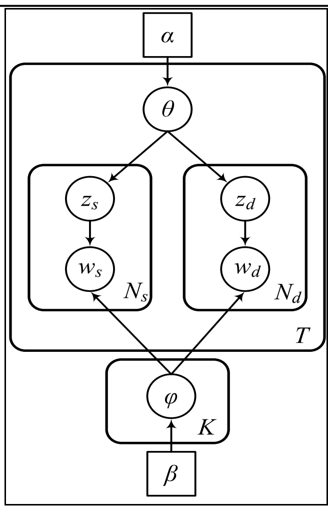

Figure 1. Generative Process and Plate Notation of JTT.

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

Table 1. List of notations.

Exact inference for JTT is intractable. Hence, we use Markov Chain Monte Carlo (MCMC) Gibbs sampling. Rao-Blackwellization (Bishop, 2006) is used to reduce sampling variance by collapsing latent variables θ and φ. Owing to space constraints, we omit the full derivation details. The full joint can be written succinctly as follows:

$$P(W^S, W^D, Z^S, Z^D) = \left(\prod_{t=1}^{T} \frac{B\left(n^S_{t(s),[\,]} + n^D_{t(d),[\,]} + \alpha\right)}{B(\alpha)}\right) = \left(\prod_{t=1}^{K} \frac{B(n^S_{k,[\,]} + n^D_{k,[\,]} + \beta)}{B(\beta)}\right) \quad (3)$$

Omission of a latter index in the count variables, denoted by [ ], corresponds to the row vector spanning over the latter index. The corresponding Gibbs conditional distributions for z^S and z^D are detailed below, where the subscript $(\neg(t,j))$ denotes the value of the expression excluding the counts of the term (t,j):

$$p(z^S_{t,j} = k | \dots) \propto \frac{\left(n^S_{t(s),k}\right)_{\neg(t,j)} + n^D_{t(d),k} + \alpha}{\sum_{k=1}^{K}\left(\left(n^S_{t(s),k}\right)_{\neg(t,j)} + n^D_{t(d),k} + \alpha\right)} \times \frac{\left(n^S_{k,v}\right)_{\neg(t,j)} + n^D_{k,v} + \beta}{\sum_{v=1}^{V}\left(\left(n^S_{k,v}\right)_{\neg(t,j)} + n^D_{k,v} + \beta\right)} \quad (4)$$

$$p(z^D_{t,j} = k | \dots) \propto \frac{n^S_{t(s),k} + \left(n^D_{t(d),k}\right)_{\neg(t,j)} + \alpha}{\sum_{k=1}^{K}\left(n^S_{t(s),k} + \left(n^D_{t(d),k}\right)_{\neg(t,j)} + \alpha\right)} \times \frac{n^S_{k,v} + \left(n^D_{k,v}\right)_{\neg(t,j)} + \beta}{\sum_{v=1}^{V}\left(n^S_{k,v} + \left(n^D_{k,v}\right)_{\neg(t,j)} + \beta\right)} \quad (5)$$

We learn our joint topic model from a random traffic-weighted sample of 10,000 transitions, which are randomly sampled from the development set outlined in Section 3.2[5]. The decision to use this sample of 10,000 transitions is based on the observation that there were no statistically significant performance gains for models trained on more than 10k transitions. The Dirichlet hyperparameters are set to $\alpha = 50/K$ and $\beta = 0.1$ according to the values suggested in (Griffiths and Steyvers, 2004). The number of topics, K, is empirically set to 50. We also conducted pilot experiments with other hyperparameter settings, larger transition sets and more topics but we found no substantial difference in the end-to-end performance. Although increasing the number of topics and modeling more volume usually results in lowering 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). For all reported experiments we use the posterior estimates of our joint model learned according to the above settings. In our discriminative interestingness model, we use three classes of features from JTT to capture the latent topic distributions of the source page, the destination page, and the joint topics for that transition. These correspond to source topic features (Z^S, labeled as JTTsrc in charts), destination topic features (Z^D, labeled as JTTdst), and transition topic features (Θ, labeled as JTTtrans). Each of these three sets comprises 50 features, for a total of 150.Θ is the distribution over joint src and dst topics that

[5] Note that we use the development set to train our semantic model since it is ultimately used to generate features for our discriminative learner of Section 4. Since the learner is trained using the training set, this strategy avoids overfitting our semantic model to the training set.

appear in a particular transition. Z^S and Z^D are the actual topic assignments for individual words in src and dst. Upon learning the JTT model, for each K topics, we get a probability of that topic appearing in the transition, in the src, and in the dst document (by taking the posterior point estimates for latent variables Θ, Z^S, Z^D respectively). The GBDT implementation we use for our discriminative model performs binning of these real-valued features over an ensemble of DTs.

5 Experiments

We evaluate our interestingness model on the task of proposing k anchors on a page that the user will find interesting (*highlighting* task). Recall the interestingness function σ from Eq. 1. In the highlighting task, a user is reading a document $s \in U$ and is interested in learning more about a set of anchors. Our goal in this task is to select k anchors that maximize the cumulative degree of interest of the user, i.e.:

$$\underset{A_s^k=(a_1,\ldots,a_k|a_i\in A_s)}{\operatorname{argmax}} \sum\nolimits_{a_i\in A_s^k} \sigma(s,a_i) \qquad (6)$$

In other words, we consider the ideal selection to consist of the k most interesting anchors according to $\sigma(s,a)$.We compare the interestingness ranking of our models against a gold standard function, σ', computed from our test set. Recall that we use the aggregate click frequency of an anchor as a proxy for its interestingness. As such, the gold standard function for the test set is computed as:

$$\sigma'(s,a) = p(a|s) \qquad (7)$$

where $p(a|s)$ is the probability of a user clicking on the anchor a when viewing s.

Given a source document s, we measure the quality of a model's interestingness ranking against the ideal ranking defined above using the standard nDCG metric (Manning et al., 2008). We use the interestingness score of the gold standard as the relevance score.

Table 2 shows the nDCG results for two baselines and a range of different feature sets. The first high-level observation is that the task is difficult, given the low baseline results. Since there are many anchors on an average page, picking a random set of anchors yields very low nDCG scores. The 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 (not unique) decreases in these sets from 170 to 94 to 41[6]. The second baseline illustrates that it is not sufficient to simply pick the top n anchors on a page.

Next, we see that using our set of anchor features (see Section 4.1) in the regression model greatly improves performance 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 differs according to the popularity of the source content, possibly also with the length of the page. Our best performing model is the one using anchor features and all three sets of latent semantic features (Table 2, row 6; source, destination, and transition topics).

The biggest improvement is obtained on the HEAD 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 features (Table 2, rows 4, 5), 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 2 further 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 hypothesize that this is because the average number of anchors in a TAIL page is less than half of that in a TORSO page.

[6] Wikipedia editors tend to spend more time on more frequently viewed documents, hence they tend contain more content and more anchors.

nDCG %	HEAD				TORSO				TAIL			
n	@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+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**	**38.01**

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

Not shown in these results are the effects of using user session features. We consistently found that these features did not improve upon the configurations where anchor and JTT features are used. We do not, however, rule out the potential of such features on this task, especially in light of our data analysis observations from Section 3.3, which suggest an effect from these factors. We leave a more in-depth study of the potential contribution of these types of features for future research.

We now address the question how our unsupervised latent semantic features perform compared to the editor-assigned categories for Wikipedia pages, for two reasons. First, it is reasonable to consider the manually assigned Wikipedia categories as a (fine-grained) oracle for topic assignments. Second, outside of Wikipedia, we do not have the luxury of manually assigned categories/topics. As illustrated in Figure 3, we found that Wikipedia categories outperform the JTT topic features, but the latter can recover about two thirds of the nDCG gain compared to Wikipedia categories.

Finally, in the HEAD part of the data, we have enough historical clickthrough data that we could directly leverage for prediction. We conducted experiments where we used the prior probability $p(d|s)$ obtained from the development data (both smoothed and unsmoothed). Following this strategy we can achieve up to 65% nDCG@10 as shown in Figure 4 where the use of prior history (labeled "History: Target | Source Prior") is compared to our best model and to baselines. As stressed before, in most real-life applications, this is not a viable option since anchors or user-interaction logs are unavailable. Even in web browsing scenarios, the TORSO and TAIL have no or only very sparse histories. Furthermore, the information is not available in a "cold start" scenario involving new and unseen pages. We also examined whether the general popularity of a target page is sufficient to predict an anchor's interestingness, and we found that this signal performs better than the baselines, but significantly worse than our models. This series is labeled "History: Target Prior" in Figure 4.

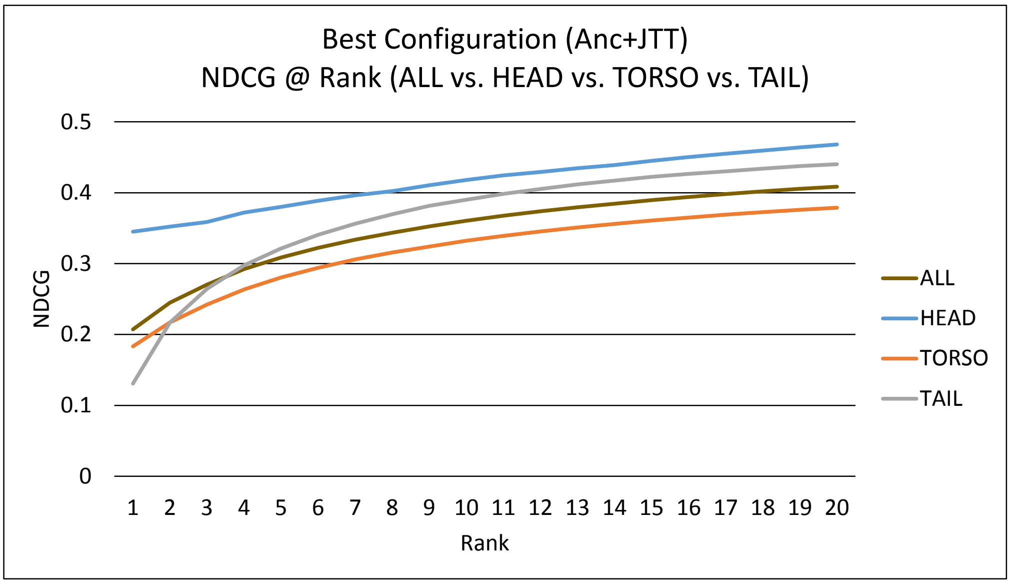

Figure 2. NDCG comparison across overall performance (ALL) versus HEAD, TORSO, and TAIL subsets, on the Highlighting task.

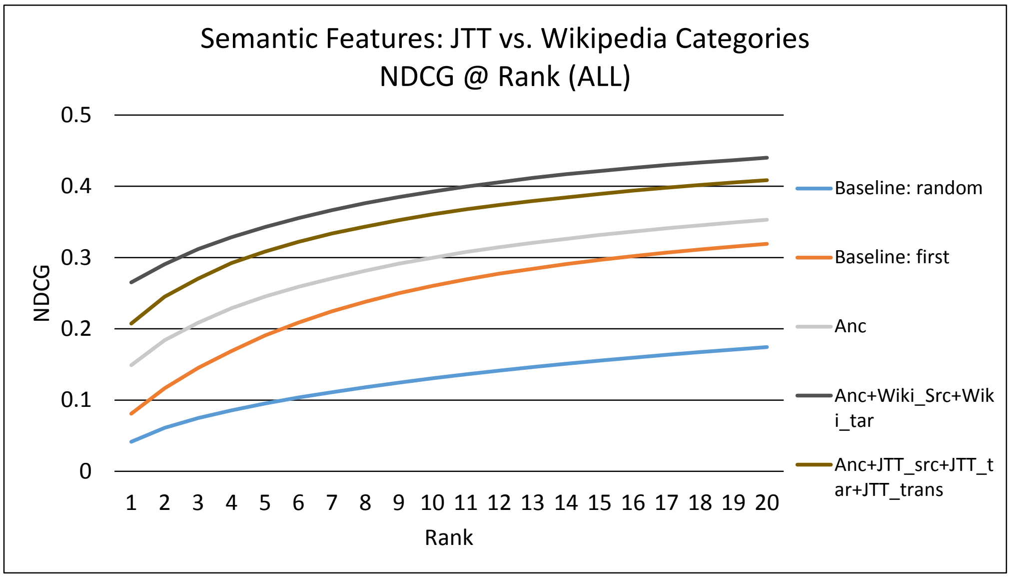

Figure 3. JTT features versus Wikipedia category features on Highlighting task.

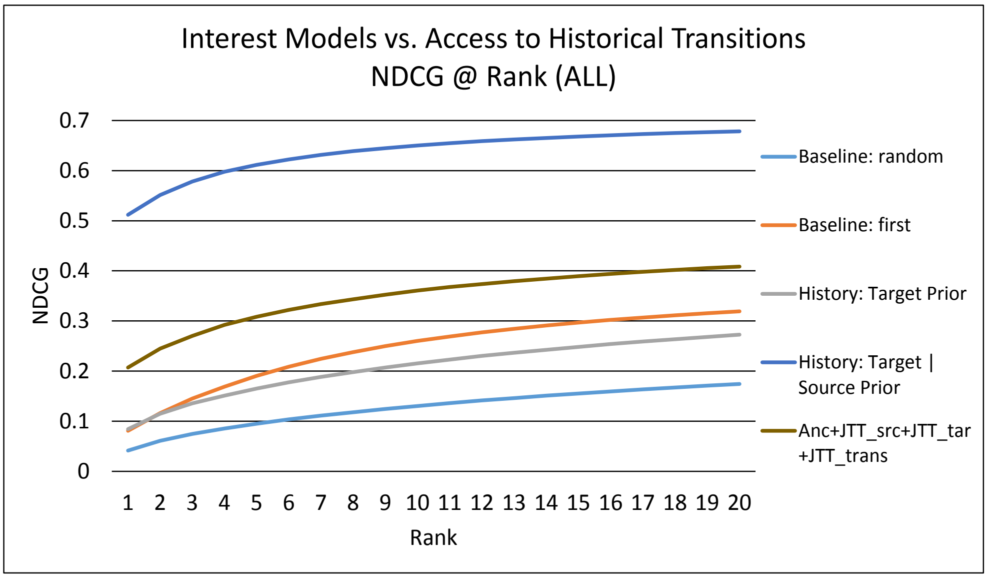

Figure 4. Highlighting task comparison between baselines, best configuration using JTT, and models with historical transitions.

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 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 τ(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 5 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.

Cumulative Precision %	HEAD				TORSO				TAIL			
n	@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+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**

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

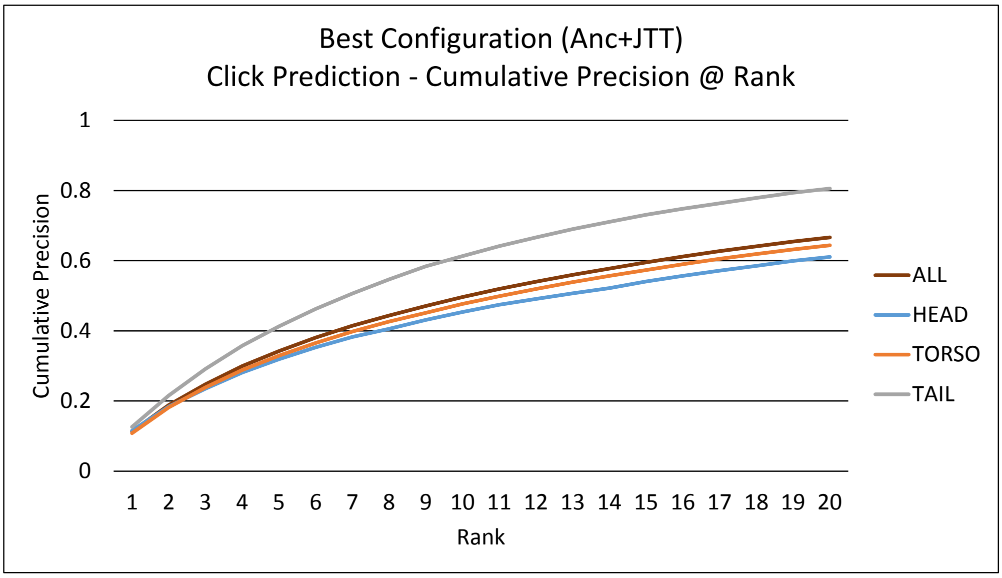

Figure 5. Overall performance (ALL) versus HEAD, TORSO, and TAIL subsets on click prediction.

6 Conclusion and Future Directions

We presented a notion of an IT on a page that is grounded in observable browsing behavior during content consumption. We implemented a model for prediction of interestingness that we trained and tested within the domain of Wikipedia. The model design is generic and not tied to our experimental choice of the Wikipedia domain and can be applied to other domains. Our model takes advantage of semantic features that we derive from a novel joint topic transition model. This semantic model takes into account the topic distributions for the 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 editor-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. First, for an application that marks interesting ITs on an arbitrary page, we would need a detector for IT candidates. A simple first approach would be to use a state-of-the-art Named Entity Recognition (NER) system to cover at least a subset of potential candidates. 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 Lin et al. 2012 and the references cited therein). 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 jointly modeling interestingness and candidate detection.

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. User session features did not contribute to model performance when used in conjunction with other feature families, but closer investigation of these features is warranted for more personalized models of interestingness. Finally, a number of options regarding JTT 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.

References

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

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

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

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

Bishop, C.M. 2006. Pattern Recognition and Machine Learning. Springer.

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. In *Proceedings of IJCAI*. pp. 2071-2077.

Claypool, M., Le, P., Wased, M., & Brown, D. 2001a. Implicit interest indicators. In *Proceedings of the 6th international conference on Intelligent user interfaces* (pp. 33-40). ACM.

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

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

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

Gamon, M., Yano, T., Song, X., Apacible, J. and Pantel, P. 2013. Identifying Salient Entities in Web Pages. In *Proceedings CIKM*. pp. 2375-2380.

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

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. In *Proceedings of ICML*. pp. 13-20.

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

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

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

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

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

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

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

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

Mueller, F., & Lockerd, A. 2001. Cheese: tracking mouse movement activity on websites, a tool for user modeling. In *CHI'01 extended abstracts on Human factors in computing systems* (pp. 279-280). ACM.

Paranjpe, D. 2009. Learning document aboutness from implicit user feedback and document structure. In *Proceedings of CIKM*. pp. 365-374.

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

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

Varela, F. J., Thompson, E. T., & Rosch, E. 1991. The embodied mind: Cognitive science and human experience. *The MIT Press*.

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