Understanding Document Aboutness Step One: Identifying Salient Entities

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

Tae Yano TAEY@CS.CMU.EDU
Carnegie Mellon University
Pittsburgh, PA, USA

Xinying Song XINSON@MICROSOFT.COM
Microsoft Research
Redmond, WA, USA

Johnson Apacible JOHNSONA@MICROSOFT.COM
Microsoft Research
Redmond, WA, USA

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

Abstract

We propose a system that determines the *salience* of entities within web documents. Many recent advances in commercial search engines leverage the identification of entities in web pages. However, for many pages, only a small subset of entities are important, or central, to the document, which can lead to degraded relevance for entity triggered experiences. We address this problem by devising a system that scores each entity on a web page according to its centrality to the page content. We propose salience classification functions that incorporate various cues from document content, web search logs, and a large web graph. To cost-effectively train the models, we introduce a novel *soft labeling* methodology that generates a set of annotations based on user behaviors observed in web search logs. We evaluate several variations of our model via a large-scale empirical study conducted over a test set, which we release publicly to the research community. We demonstrate that our methods significantly outperform competitive baselines and the previous state of the art, while keeping the human annotation cost to a minimum.

1. Introduction

The concept of *Salience* or *Aboutness* has been investigated in many fields of research, from linguistics to semiotics, and from sociology to psychology. While dictionary definitions look deceptively simple ("*most noticeable or important*" (OED), "*state or condition of being prominent*" (Wikipedia)), the notion of salience is very hard to pin down in practice. A number of observations around salience might be uncontroversial: (1) Salience on the one hand and relevance or importance on the other are not the same. An entity or notion A in a text can be highly salient, yet unimportant to the reader: maybe the reader is already very familiar with A, or they simply do not care about anything having to do with A; (2)

Salience is a function of the structure of a text, and indirectly a function of the intention of the author, as opposed to a function of the reader's intent or needs.

In the context of the Web, salience has very practical implications, for instance in connecting people, entities, and content in a "Web of Things" paradigm (Dalvi, Kumar, Pang, Ramakrishnan, Tomkins, Bohannon, Keerthi, & Merugu, 2009). For example, if your friend has liked a review about a recently released movie, Bing will annotate your search result page with your friend's photo when you issue a query for that movie name. If you are browsing the Baltimore Ravens page on Yahoo! Sports, content related to the team is recommended. For an entity search, Google displays pertinent facts from its Knowledge Graph in the sidebar, including relationships with other entities.

Not all of these links can be extracted with equal precision. Links between people and entities can be reliably acquired through social signals such as *likes* or *+1s* (Pantel, Gamon, Alonso, & Haas, 2012; Muralidharan, Gyongyi, & Chi, 2012). Links between people and content can be mined from web search logs (Agichtein, Brill, & Dumais, 2006; Baeza-Yates, 2004). Whereas entity pages from popular or wrapped sites create links between entities and content, harvesting them from arbitrary web pages has proven difficult.

One of the reasons for this difficulty is that web pages contain a large number of non-salient entities. On average we found 66 entities per page on a random sample of crawled web pages at a commercial search engine. Named Entity Recognition systems, such as (Finkel, Grenager, & Manning, 2005; Cucerzan, 2007), are generally effective at identifying entities on web pages[1]. However, in our samples we found that fewer than 5% of these entities are salient to the web page. Consider, for example, how many entities may be mentioned in a news story about the Syrian civil war, and how few of them are central to what the story is about. Qatar may be mentioned as a location where a meeting between Syrian opposition groups took place, and the Associated Press may be credited as the source of the story. "Qatar" and "The Associated Press" are peripheral to the story, i.e., not salient. In contrast, "Syria" and the name of the central opposition fighters in the story have high salience. Without a notion of salience, many irrelevant links would be established between entities and documents. In a social search experience such as the one presented on Bing, if your friend has "liked" this article and you queried for *The Associated Press*, it would be detrimental to the experience to present a social annotation with your friend's "like".

Knowing the salient entities in a web page not only enables us to focus on the relevant edges between people and entities, it can also add edges from content to content, pivoting on entities and enabling pivot recommendations. In addition, new experiences are made possible such as stamping web documents with detailed entity cards for its salient entities. One might also expect gains in traditional search by incorporating the entities as relevance features and by improving query-independent summaries via targeted sentence selection based on key entities.

User interactions with search engines and web content provide a unique opportunity for modeling entity salience. By observing search usage behavior at a commercial search engine, we can collect links between the things people search for and the documents that satisfy their needs. We demonstrate that this signal can serve as soft supervision for learning a salience model.

1. Performance of NER systems vary on the entity type, but high accuracy is often observed for classes such as people, companies, and places.

We propose scalable models for learning to rank and classify entities according to their salience to a document. We leverage web search logs to automatically acquire *soft labels* as a supervision signal for our training data. We train our models on a large number of web pages, leveraging features from document content, page classifiers, and a web graph. Finally we show empirical evidence, on data representing the `HEAD` and `TAIL` distributions of the web, that our methods significantly outperform the previous state of the art on various ranking and classification metrics. As this is the first dataset created of its kind, we release it publicly to the research community.

The main contributions of this paper are:

- We devise a notion of entity salience and frame the problem of understanding the aboutness of a document as determining the most salient entities in the document;
- We model the task of entity salience detection as a weakly supervised machine learned model, generating labeled training data via usage behaviors found in web search logs;
- We present empirical evidence that our system significantly outperforms previously established baselines; and
- We publicly release test sets consisting of URLs and their entity mentions randomly drawn from head and tail distributions in a commercial web search engine along with gold standard salience judgments.

2. Related Work

Understanding the meaning or aboutness of a document has received attention from both a theoretical (Putnam, 1958; Hutchins, 1977; Bruza, Song, & Wong, 2000) and practical perspective. In the latter approaches, driven by application-specific demands, computational models have decomposed aboutness and focused on detecting aspects of aboutness such as key terms (Yih, Goodman, & Carvalho, 2006; Irmak, Brzeski, & Kraft, 2009; Paranjpe, 2009), latent semantic spaces/topics (Landauer & Dumais, 1997; Blei, Ng, & Jordan, 2003), and summaries (Salton, Allan, & Buckley, 1993; Kupiec, Pedersen, & Chen, 1995; Hovy & Lin, 1998). Motivated by our entity-centric web experiences outlined in Section 1, we believe that an important aspect of document aboutness is *salient entities*, i.e., those entities in a document that are central to its meaning.

Most related to our work, and subject to a direct comparison throughout our paper, is the current state of the art described in Paranjpe (Paranjpe, 2009) where the focus is on the detection of key terms in web pages. Web search log information is used for soft labeling of term salience that then serves as training data for a supervised salience scoring function. Although our system also uses auto-labeling and supervised training techniques, there are three key differences between our proposed method and theirs. First, our notion of document aboutness is entity-centric, i.e., we consider the identification of salient entities, as opposed to salient terms of any kind. Second, our soft labeling method is different from theirs and we demonstrate that it outperforms it and is more robust to effects of popularity and presentation order of URLs in the SERP. Finally, our feature set is a significant extension of the set of features Paranjpe utilizes.

The keyword extraction task can be seen as related to entity salience, where keywords and key phrases are a superset of salient entities in a document. Keyword extraction is often

addressed in the context of various document understanding tasks, most often in extraction based summarization or abstract generation (Hulth, 2003; Marcu, 1997; Hovy & Lin, 1998), and more recently in (online) contextual advertisement or keyword appraisal (Yih et al., 2006). Linguistic cues for keyword extraction are investigated in (Hulth, 2003), for example, who conducts empirical tests on the informativeness of shallow syntactic cues in their key phrase extraction system. In his seminal work, (Marcu, 1997) illustrates the importance of discourse context in assessing term centrality in documents. Others (Hutchins, 1977; Hjørland, 2001; Bruza et al., 2000) examined various syntactic, semantic, and discourse properties that indicate centrality and focus (which are closely related to aboutness). This research also explored how those insights can be applied to practical aboutness-related tasks such as topicality classification, subject indexing, but not entity salience. Our proposed system does not use any features based on linguistic analysis due to the running time complexity of most analyses that would in turn make it difficult to process a large crawl of the web, but it should be clear that it would be possible to incorporate such features in our learning framework.

A second approach to keyword extraction is purely based on term statistics, without recourse to more complex linguistic structure. For example, term frequency statistics and various term weighting schemes are commonly used to score the specificity, or importance, of a term in information retrieval (Manning, Raghavan, & Schütze, 2008, Chapter 7). The same idea motivates the vector space model and its application in ad hoc document retrieval (Salton & Buckley, 1988). Salton applied the vector space model to document indexing and key phrase generation (Salton, Wong, & Yang, 1975). We build on this insight by incorporating a set of term and document frequency features in our models, but also extending it to other signal sources such as the web graph and search logs.

We use supervised machine learning to build our models of entity salience, a method that has been used widely for various tasks in web document processing. Machine learning offers a principled way to calibrate signals from heterogeneous sources, which is crucial when incorporating diverse (e.g. document content, term-weighting, web graph) insights into one system. The drawback of supervised learning is the cost of gathering "supervision", or annotation for training data. This is especially problematic for newer domains or novel tasks where annotated resources have to be created from scratch. A wide variety of tactics are employed in the literature to overcome this bottleneck, for a theoretical perspective on these approaches see (Zhu, 2005). One well studied approach to obtain relevance-related supervision for web document training data is the use of web search logs: the click behavior that is recorded in these logs can serve as implicit user feedback and hence indicate relevance of a document to a user. This signal has been exploited as a surrogate for relevance annotation in document retrieval systems (Joachims, 2002; Joachims, Granka, Pang, Hembrooke, & Gay, 2005; Holland, Ester, & Kießling, 2003; Radlinski & Joachims, 2005). Recently, these implicit annotations have been used for other web document tasks beyond retrieval. In (Xu, Yang, & Li, 2009), the authors used web search logs to provide guidance for generative entity mining and a clustering model. Komachi and Suzuki (Komachi & Suzuki, 2008) designed an automatic annotation scheme on the basis of a web search log for a semantic relation extraction system. Paşca and van Durme (Paşca & Durme, 2007) use a web search log to obtain weak supervision for their entity attribute discovery. Irmak et al. (Irmak et al., 2009) used statistics from clicks on keywords in Yahoo! Shortcuts to train

their machine learned model. In our system, we exploit web search logs by designing a soft labeling function for entity salience that is based on user behavior information in the logs. We demonstrate that this function successfully approximates entity salience and hence can be used as a supervision signal, removing the need for manual salience annotation.

3. Scoping

In this section, we begin with a functional definition of entity salience and then describe our document and entity landscape via a manual investigation of random web pages. We then test the hypothesis that a state-of-the-art NER system, although insufficient for identifying salience, can in fact be an effective candidate generator of salient entities. Finally, we consider the question of whether or not assigning salience to entities is a trivial process.

3.1 Entity Salience

What constitutes an entity has been cause of many philosophical debates. For our purposes, we consider something an entity if it has or reasonably could have a Wikipedia page associated with it. This would include people, places, companies, as well as events, concepts, and famous dates.

We consider the following working assumptions in building our entity salience ranking system:

- **Local scoping:** The salience of an entity can be solely determined by how the entity is presented within the document. In other words, entity salience can be effectively computed from the local context, or what is available in the document itself.
- **Invariable perception:** Entity salience can be assessed independently from the intentions or interests of its users/readers, and independently from the prior importance of the entity as it exists outside of the document.

Entity salience is distinct from two other aspects of aboutness: *entity importance* and *entity relevance*. The importance of an entity refers to its influence or substantiveness outside of the scope of the document. For example, although *Barack Obama* is a very important entity, he can be peripheral to some news stories. On the other hand, the relevance of an entity is inherently subjective to the reader's perspective and intent. For example, in an article about an expressionist art tour featuring Munch's *The Scream*, a reader's hometown would be perceived as much more relevant to her than to a non-resident reader.

Although local scoping suggests that the evidence for entity salience can be derived most effectively from the document content, it is important to note that extra-document information such as incoming anchor links and user clickthrough data provide important signal, and will be leveraged by our models. Also, by assuming the source of salience to be local to a document, we limit the search space to those entities in the document.

3.2 Qualitative Study

We conducted a small manual inspection of web pages in order to get a first perspective at the difficulty and scope of our problem. We sampled 50 documents, randomly chosen from

a traffic-weighted sample of documents from a commercial web search index. We examined the content of the pages in a web browser and made a list of all entities and the salient ones.

The largest classes of documents were narrative stories, such as news and blog articles, YouTube pages, and entity pages (e.g., company and product pages). On average, fewer than 5% of the entities in each document were deemed salient. We observed certain cues when identifying salient entities. Unsurprisingly, salient entities tend to be mentioned in the title, headings, and/or first paragraph, and are frequently mentioned.

In the next subsections, we address two questions. First, we determine whether or not an NER system can reliably generate a set of candidate entities that contains the true salient entities on a page. Second, we discuss the complexity of determining entity salience with respect to the cues we identified in our study.

3.2.1 NER as a Candidate Generator

By our local scoping assumption, any salient entity is contained in its document. Hence, a system that is capable of identifying each entity in a document would serve as a candidate generator for a salience ranking system. Below we test whether an NER system can indeed generate true salient entities in its list of entities. In this paper, we use a proprietary state-of-the-art NER system, trained using the perceptron algorithm (Collins, 2002).

After the annotation exercise, we ran our NER engine on the content of the web documents. We then compared for each page the set of automatically identified entities to the human annotation to ensure that the truly salient entities are in fact mostly recognized by the NER engine.

We found that in 91% of the documents, at least one of the salient entities is in the candidate entity set identified by our NER tagger. For over 90% of these pages, all the human annotated salient entities are captured by our NER engine. Therefore it is reasonable to use the NER system as a candidate generator.

3.2.2 Is the problem trivial?

Most entities in our sampled data are not salient: on average only 4.4%. Below we address whether simple cues for entity salience are so straightforward that a heuristic would suffice to identify them.

We observed many cases where cues were not reliable or conflicted with each other, making heuristic design a difficult proposition. For example, the presence of an entity in a title string is often a good indicator for salience. However, being included in the title (or in the first paragraph) is neither a necessary nor a sufficient condition for salience. Authors may omit the most salient entities from titles or first paragraphs when these entities are obvious for the audience, or when building up a story rather than cutting straight to the point. For example, we noted a news article about Hurricane Sandy in the *Local* section of the New York Times that was simply titled *Cold, Dark and Damp*. Based on these observations, we believe that a machine learned model that can combine evidence from a multitude of signals is a better approach than developing simple heuristics. In the following section, we describe our models that leverage signal not only from the document content, but also its domain, a web graph, and web search logs. The experimental results in this

paper show that using multiple sources of evidence in conjunction significantly outperforms the use of simple cues.

4. Model

We set up our learning task as a supervised learning scenario. Below we first formally define our task, then describe methods to automatically generate labeled training data, and finally present our learning algorithm and features.

4.1 Task Definition

Let $\mathbf{D}$ and $\mathbf{E}$ be the sets of all documents and entities on the web, respectively. Let $\mathbf{E_d} \subset \mathbf{E}$ be the set of entities mentioned in $d \in \mathbf{D}$. We formally define the aboutness task as learning the function:

$$\sigma : \mathbf{D} \times \mathbf{E} \to \mathbb{R} \tag{1}$$

where $\sigma(d, e)$ reflects the salience of e in d[2].

We denote the ranking of $\mathbf{E_d}$ according to σ as:

$$\mathbf{R_d^S} = (e_1, ..., e_{|E_d|} \mid e_i \in E_d, \sigma(d, e_i) \geq \sigma(d, e_{i+1}))$$

where pairs of entities with tied scores are ordered randomly. We define the ranking function

$$R_\sigma : \mathbf{D} \times \mathbf{E} \to \mathbb{N} \tag{2}$$

such that $R_\sigma(d, e)$ equals the rank of e in $\mathbf{R_d^S}$[3].

4.2 Soft Labeling

Manually labeled salience judgment data is hard to obtain in sufficient quantity to train a model for diverse web sites in the head and tail distributions of URLs. Instead of manually labeled data we rely on a *soft labeling* approach that uses behavioral signals from web users as a proxy for salience annotation. Mining a web search log from a commercial search engine, we can observe how users query for entities and which URLs they subsequently click on. Individual clicks indicate a user's interest in a URL based on their entity query, i.e., they indicate the relevance of the entity in the URL to the user. In aggregate, the combined interests for an entity/URL pair will correlate with the entity being salient, since users are less likely to search for an entity and then examine a page that is not about that entity. This "soft label" is available for pages that receive enough traffic to derive reliable user click statistics. In other words, we obtain this signal predominantly for queries and URLs in the head of the distribution. However, while this is true for the supervision signal, the learned model uses features that are independent of user behavior, and hence it can generalize to the tail of the distribution.

2. We fix $\sigma(d, e) = 0$ for all $e \notin \mathbf{E_d}$.
3. $R_\sigma(d, e)$ is not defined for $e \notin \mathbf{E_d}$.

4.2.1 Click Attractivity

A simple click measure is Clickthrough Rate (CTR), i.e. the rate at which users click on a URL given a query. Paranjpe (Paranjpe, 2009) points out, however, that CTR is very much biased towards the top-ranked result on the SERP which tends to receive the bulk of user clicks. Instead, they propose to use Click Attractivity (CA) as a search log based metric that correlates with salience. CA for a term t and document d is defined as:

$$CA(t,d) = \frac{clicks(t,d)}{clicks(t,d) \ + \ skips(t,d)} \tag{3}$$

where $clicks(t,d)$ is the number of times users clicked on d for a query containing t, and $skips(t,d)$ is the number of times users clicked on another document d' that is ranked at a lower position than d, where d is in the top-5 results. Both *clicks* and *skips* are aggregated over all queries that contain t and lead to at least 32 instances where document d is displayed in the search result page (SERP). In its original setting, CA was used for any term t in a document. For this paper, only terms that are entities are considered, i.e., $t = e$.

Upon closer inspection, CA has its own problems, however. It is based on the assumption that when a user looks for entity e, the URL she clicks on is the one that is most *about* e compared to the other URLs in the SERP. That assumption does not hold (especially for the low salience signal) in several common circumstances:

1. *Recency trumps salience*: Assume that entity e is involved in some recent gossip news. The user will be most interested in the latest gossip about e (which provides a good signal for salience) but will hardly ever click on the IMDB or Wikipedia page for e, although on these pages e is very salient.

2. *Popularity trumps salience*: Within a set of URLs that are equally about e, some of the sites might be more popular than others (e.g., a celebrity home page will be more popular than a page about her maintained by a fan.) This will distort the CA score.

3. Although the claim that CTR is subject to position bias is true, we argue that CA is similarly affected. If the user is generally more likely to click on a URL in the top position, this also means that she is less likely to skip that top position and hence that CA is also influenced by position bias.

In summary, we believe that CA is influenced by many factors that are not related to salience.

4.2.2 Entity Query Ratio

To circumvent the pitfalls associated with both CTR and CA, we propose a different soft labeling function that sidesteps the issues of position bias, popularity and recency by aggregating over only the queries that lead to clicks on a URL without taking the number of views (CTR) or the number of skips (CA) into account. This function is based on the simple assumption that a page that is about entity e will get most of its clicks from queries about entity e. We define Entity Query Ratio (EQR) for entity e and document d by looking at all queries that lead to a click on d. Within that set of queries, we calculate the ratio of the number of clicks from queries containing e to the number of clicks from all queries. The

notion of a query *containing* an entity e can be defined in two ways. The more restrictive notion is that the query and the entity are an exact match, and the softer notion is that the query contains the entity, but may also contain other words. We experimented with both but found that the *exact match* method consistently performs best for soft labeling, so we hereon restrict our discussion to the latter. The formal definition of EQR is:

$$EQR(e,d) = \frac{clicks(e,d)}{\sum_{q \in Q} clicks(q,d)} \tag{4}$$

where Q is the set of all queries and $clicks(e,d)$ is redefined as the number of times users clicked on d for a query matching e.

4.3 Learning Algorithm

Both CA and EQR soft labels produce a continuous score between 0 and 1. There are at least three different possibilities for modeling the prediction of the score. The most natural solution for prediction of a continuous salience score is regression, i.e., a model that tries to fit a curve of predicted salience scores to the curve of observed (via soft label) salience. The modeling can also be cast as a ranking problem where the model's task is to rank the top n most salient entities in a page in the correct order. Finally, one could map the regression task into a binary classification task where each entity above a soft label threshold τ is considered salient and otherwise non-salient. This approach faces a number of difficulties, however: the best τ needs to be determined and forcing a binary decision on a continuous label is typically not very successful compared to a regression approach. We confirmed this suspicion with several classification experiments and for the remainder of the paper we will only consider regression and ranking as appropriate learning tasks.

For both ranking and regression, we utilize boosted decision trees (Friedman, 1999) as our learning algorithm system. This algorithm has a number of advantages: it has been widely used and yields high accuracy; it does not require feature normalization; it can handle a mix of real-valued and binary features; and it can capture non-linearities between features. The hyperparameters are the number of iterations, learning rate, minimum instances in leaf nodes, and the number of leaves. The particular settings for our experiments are described in Section 6.1.

4.4 Features

We represent each entity/document pair $\langle e, d \rangle$ as a vector of features, listed in full in Table 1. At the highest level, there are three distinct classes of features: (1) those that are computed from properties of e and the whole document collection D, labeled $\mathrm{F}_{e,\mathbf{D}}$; (2) those that are solely computed from properties of d, labeled F_d; and (3) those that are computed from properties of e in d, labeled $\mathrm{F}_{e,d}$. Document features, F_d, further sub-divide into categorical features representing the page classification of d, features of the document URL, and length features. Entity/document features, $\mathrm{F}_{e,d}$, are subcategorized into structural features that relate e to the structure of d, web graph features that indicate the frequency of e in inlinks and outlinks, position features that capture the location of e in d, and finally features that capture the frequency of e in 17 different page segments that are automatically identified based on visual properties (see (Cai, Yu, Wen, & Ma, 2003) for more details and (Song,

Class	Family	Description
$F_{e,\mathbf{D}}$	Corpus features	tf, df, idf, tf.idf of e in $\mathbf{D}$
		same in $\mathbf{D_2}$ where $\mathbf{D_2}$ contains the documents in the top domain of d
F_d	Page classification	**page category of d**
		junk page score of d
		inlink spam page score
		spam page confidence
	Url	**top level domain**
		Url depth
	Length	length of d
		length of main content of d
$F_{e,d}$	Structural	**e is in the title**
		e is in a meta key word
		e is in the visual title
		e is in bold font
		e is in emphasized font
		e is in italics
		norm freq of e in table headers
		norm freq of e in table body
		norm freq of e in list body
		norm freq of e in text
		norm freq of e in title
		norm freq of e in meta key words
		norm freq of e in visual title
		norm freq of e in bold font
		norm freq of e in emphasized font
		norm freq of e in italics
		e is part of Url
		e is in domain part of Url
		e is in last part of Url
	Web graph	norm freq of e in out anchors
		norm freq of e in in anchors
	Position	first offset of e
		last offset of e
		mean offset of e
		standard deviation of offset of e
		first offset of e in main content
		last offset of e in main content
		mean offset of e in main content
		standard deviation of offset of e in main content
	Page segmentation	**norm freq of e in several different page segments**

Table 1: All features, classified according to class and family, used to train our salience models. Boldface features are new compared to Paranjpe (Paranjpe, 2009).

Liu, Wen, & Ma, 2004) for other work that used visual blocks as input). Boldface features in Table 1 indicate additions to the feature set used in (Paranjpe, 2009). The total number of resulting features is 7,544, consisting of 59 numeric features and 7,485 binary features representing all observed values of the categorical features.

5. Evaluation Methodology

In this section, we describe how we construct our test datasets and the metrics that we use for analyzing the performance of a salience function σ.

5.1 Test Set Construction

Let ρ be a graded relevance scoring function for a document d and entity e:

$$\rho : \mathbf{D} \times \mathbf{E} \rightarrow \{\mathbf{MS}, \mathbf{LS}, \mathbf{NS}\} \tag{5}$$

where for $\rho(d, e)$:

- **Most Salient (MS)** indicates that d is mostly about e, or e plays a prominent role in the content of d;
- **Less Salient (LS)** indicates that e plays an important role in some parts of the content of d; and
- **Not Salient (NS)** indicates that d is not about e.

We define a test set $\mathbf{T} = \{\rho, \mathbf{\Delta}\}$ where $\mathbf{\Delta} = \{(d, \mathbf{E_d}) : d \in \mathbf{D}, \mathbf{E_d} \subset \mathbf{E}\}$ as a collection of pairs of web pages and entities for which we have a gold standard ρ.

We start by constructing a universe of web pages by mining all the shared URLs on the full firehose of Twitter.com during May 2012. We discarded any URL that redirected to a query on a search engine. Second, we removed all links to YouTube.com since we observed during our study in Section 3.2 that the salient entities here are most often trivial to identify using a simple heuristic over the DOM tree. Finally, we removed any URL that did not receive at least three clicks on a commercial search engine during a six month period overlapping with our Twitter data. The final set consists of over half a million URLs, for which we have access to a full crawl of the content.

From this set of web pages, we produce 2414 manually annotated test cases for our experiments, spanning two test sets outlined below. Each test set consists of randomly sampled web pages such that each page contains fifty or fewer entities to facilitate manual annotation. The first set, labeled `HEAD`, consists of a traffic-weighted random sample of web pages from our universe of URLs, where the traffic weights are estimated using the number of clicks each URL received during a six month period. This set represents the head distribution of our URLs. The second test set, labeled `TAIL`, consists of a uniform random sample of web pages from our universe of URLs. This set represents the long tail of the web.

For each web page in our test sets, we built the set of entity mentions by running the Named Entity Recognizer, described in Section 3.2.1, on the content of each page. There are 1228 candidate entities in the `HEAD` set and 1186 in the `TAIL` set.

$$\phi_{bin}(d,e) = \begin{cases} 1 & \text{if } \rho(d,e) = \textbf{MS} \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_{tri}(d,e) = \begin{cases} 1 & \text{if } \rho(d,e) = \textbf{MS} \\ 0.5 & \text{if } \rho(d,e) = \textbf{LS} \\ 0 & \text{if } \rho(d,e) = \textbf{NS} \end{cases}$$

Table 2: Numeric conversion functions ϕ of graded relevance scores ρ.

To complete HEAD and TAIL, we construct gold standard relevance assessments, ρ, for each entity-document pair. We used a crowdsourcing tool to collect relevance judgments (**MS**, **LS**, or **NS**, see Section 5.1) from non-expert paid judges. For each entity-document pair, we requested five judgments. We removed all judgments from *bad* judges, which were identified as those whose mean judgment score was further than two standard deviations from the mean of all judges. This resulted in the removal of four judges for HEAD and seven for TAIL. The task had fair agreement for both test sets, with a Fleiss' κ score of 0.29 on HEAD and 0.25 on TAIL. Three expert judges then adjudicated the majority vote for each entity-document pair.

The HEAD and TAIL test sets along with their gold standard annotations are available at http://research.microsoft.com/research/downloads/details/5a2ddfde-83f7-4962-9ad7-d80cd5098f38/details.aspx.

5.2 Performance Metrics

To assess the quality of a salience function σ on a test set $\mathbf{T}$, we compute the aggregate performance against the salience judgements given by the human judges. We consider two types of applications. First, rank-sensitive applications, such as those deriving relevance features for a search ranking function, require the top-K most salient entities. For these, classic IR metrics such as **nDCG** (normalized discounted cumulative gain) and **MAP** (mean average precision) are applicable (Manning et al., 2008). Second, in class-sensitive applications, such as highlighting the salient entities on a document, we require all the salient entities on the page. For this class of applications, **Precision**, **Recall**, and **F1** metrics are applicable.

Below we define **nDCG** and **MAP** with respect to a system σ, its corresponding ranking function R_σ (Eq. 2), and test set $\mathbf{T}$.

$$\textbf{nDCG}_{\mathbf{T}}(\sigma) = \frac{1}{\mid \mathbf{T} \mid} \times \sum_{(d,\mathbf{R}^\sigma_{\mathbf{d}}) \in \mathbf{T}} \frac{\sum_{r=1}^{|\mathbf{R}^\sigma_{\mathbf{d}}|} \frac{2^{\phi_{tri}(d,e_r)}-1}{\log_2(r+1)}}{IDCG(d, \mathbf{E_d})}$$

where $\phi_{tri}(d, e_r)$ maps the relevance score of e_r in d to a real-valued score according to Table 2, and $IDCG(d, \mathbf{E_d})$ is the ideal DCG if $\mathbf{E_d}$ was perfectly ranked.

$$\textbf{MAP}_{\mathbf{T}}(\sigma) = \frac{1}{\mid \mathbf{T} \mid} \times \sum_{(d,\mathbf{R}^\sigma_{\mathbf{d}}) \in \mathbf{T}} \frac{\sum_{r=1}^{|\mathbf{R}^\sigma_{\mathbf{d}}|} \phi_{bin}(d,e_r) Prec(\mathbf{R}^\sigma_{\mathbf{d}}[\mathbf{1},\mathbf{r}], d)}{\sum_{r=1}^{|\mathbf{R}^\sigma_{\mathbf{d}}|} \phi_{bin}(d,e_r)}$$

where $\mathbf{R}^\sigma_{\mathbf{d}}[\mathbf{1},\mathbf{r}] = \{e_1, ..., e_r | e_i \in R^\sigma_d\}$, $\phi_{bin}(d, e_r)$ indicates if the entity at rank r is salient or not in d (see Table 2), and:

	HEAD				
	nDCG@1	**nDCG@5**	**MAP@1**	**MAP@5**	**F1**
$\mathbf{CA_{base}}$	0.49	0.54	0.28	0.33	0.55
$\mathbf{EQR_{base}}$	0.51	0.54	0.20	0.28	0.55
CA_TFIDF	0.66	0.73	0.38	0.46	0.63
CA_PJP	0.70	0.80	0.42	0.51	0.66
CA_ALL	$0.80^{\dagger}$	$\mathbf{0.85}^{\dagger}$	$0.52^{\dagger}$	$0.57^{\dagger}$	0.70
CA_ALL_RANK	$0.80^{\dagger}$	$\mathbf{0.85}^{\ddagger}$	$0.52^{\dagger}$	0.57	0.70
EQR_TFIDF	0.60	0.71	0.32	0.43	0.59
EQR_PJP	$\mathbf{0.82}^{\ddagger}$	0.81	$\mathbf{0.54}^{\ddagger}$	$0.56^{\dagger}$	0.69
EQR_ALL	$\mathbf{0.82}^{\ddagger}$	$\mathbf{0.85}^{\dagger}$	$\mathbf{0.54}^{\ddagger}$	$\mathbf{0.58}^{\dagger}$	$\mathbf{0.75}^{\dagger}$
EQR_ALL_RANK	$0.80^{\dagger}$	$0.84^{\dagger}$	$0.52^{\dagger}$	$\mathbf{0.58}^{\dagger}$	$\mathbf{0.75}^{\dagger}$
	TAIL				
	nDCG@1	**nDCG@5**	**MAP@1**	**MAP@5**	**F1**
$\mathbf{CA_{base}}$	0.43	0.46	0.27	0.32	0.42
$\mathbf{EQR_{base}}$	0.43	0.46	0.12	0.27	0.42
CA_TFIDF	0.54	0.57	0.29	0.38	0.48
CA_PJP	0.60	0.65	$0.35^{\dagger}$	0.47	0.55
CA_ALL	$0.65^{\dagger}$	$0.76^{\ddagger}$	$0.40^{\dagger}$	$0.54^{\dagger}$	$0.61^{\dagger}$
CA_ALL_RANK	$0.73^{\ddagger}$	$0.72^{\dagger}$	$0.48^{\dagger}$	$0.54^{\dagger}$	$0.59^{\dagger}$
EQR_TFIDF	0.56	0.58	0.31	0.40	0.49
EQR_PJP	$0.65^{\dagger}$	$0.66^{\dagger}$	0.40	0.46	0.56
EQR_ALL	$0.73^{\ddagger}$	$\mathbf{0.77}^{\ddagger}$	$0.48^{\ddagger}$	$\mathbf{0.58}^{\ddagger}$	$\mathbf{0.64}^{\ddagger}$
EQR_ALL_RANK	$\mathbf{0.77}^{\ddagger}$	$0.74^{\dagger}$	$\mathbf{0.52}^{\ddagger}$	$0.56^{\dagger}$	$0.59^{\dagger}$

Table 3: Results on `HEAD` and `TAIL` using rank-sensitive metrics (nDCG and MAP) and classification-sensitive metric F1. † indicates statistical significance over the soft labeling baselines and the `tf.idf` feature configuration; ‡ further indicates statistical significance over CA_PJP (significance assessed using Student's t-Test with p-value = 0.1). Bold indicates the highest achieved score on each metric.

$$\mathbf{Prec}(\mathbf{R}, d) = \frac{\sum_{r=1}^{|\mathbf{R}|} \phi_{bin}(d, e_r)}{|\ \mathbf{R}\ |}$$

Recall and **F1** follow trivially.

6. Experimental Results

6.1 Experimental Setup

We construct two training sets, one using the Click Attractivity (CA) method from Section 4.2.1, and the other using our Entity Query Ratio (EQR) method from Section 4.2.2. We first ran our NER system on the content in our Web Page Data, discarding those pages in our `HEAD` and `TAIL` test sets, and associated all queries from the US English market of a search engine that led to a click on the pages during a six month period. We computed the CA and EQR scores for each entity. Many entity-URL pairs receive a zero score because no query mentioning the entity leads to any click on the URL. Although such an entity-URL

pair could in fact be salient (even with six months of web search log data, there is sparsity in the tail), in most cases the pair is non-salient. In our experiments, we tried configurations that included all zero-scoring entity-URL pairs, none of them, and balancing the number of zero-scoring pairs to be equal to the number of non-zero-scoring pairs via random sampling. The balanced configurations consistently and by a large margin outperformed the others, and hereon we consider only balanced configurations. For the EQR soft label, our final training set contains 66,055 entity-URL pairs; for the CA soft label the number of entity-URL pairs in the training set is 48,759[4].

To complete the training sets, for each soft-labeled entity-URL pair, we computed the features listed in Table 1. We used a commercial search engine to compute features that require web graph data or page classification. To set the hyperparameters of our regression and ranking models from Section 4.3, we perform a sweep of 144 combinations of parameter settings on a three-fold cross validation, for each system configuration.

Each system that we train and evaluate consists of three choices: soft labeling method (CA vs. EQR), feature set (selected from those listed in Table 1), and model type (regression and ranking).

We consider the following five baselines against which to test our systems:

- **CA$_{\mathbf{base}}$** and **EQR$_{\mathbf{base}}$**: The systems that use the CA and EQR soft labels as their prediction (without a learned model);
- **CA_TFIDF** and **EQR_TFIDF**: Regression models using only the `tf`, `df`, `tf.idf` features.
- **CA_PJP**: Current state-of-the-art model (Paranjpe, 2009).

We report our results on the following system configurations:

- **EQR_PJP**: Regression model with the feature set from (Paranjpe, 2009) with our soft labeling function.
- **CA_ALL** and **EQR_ALL**: Regression models with all features from Table 1.
- **CA_ALL_RANK** and **EQR_ALL_RANK**: Ranking models with all features.

6.2 System Comparison

Table 3 lists the performance of our baseline and system configurations on both the `HEAD` and `TAIL` datasets. We report nDCG and MAP scores (at 1 and 5) and F1. **EQR_ALL** and **EQR_ALL_RANK**, our best configurations, significantly outperform the soft labeling baselines, on both `HEAD` and `TAIL`, by 37% and 51% on F1, respectively. On `TAIL`, we improve on the previous state of the art, **CA_PJP**, significantly on all metrics, by 16% on F1. On `HEAD`, we show significant improvement over **CA_PJP** in the first position on both nDCG and MAP.

In general, the `HEAD` is "easier" than the `TAIL` in the sense that absolute metrics are higher, and that the choice of feature sets and soft labeling function matters less. This is not surprising for two reasons: (1) the soft label signal is reliable only in the head because

4. This discrepancy in number of training cases is due to the fact that we only compute the CA label for documents in the top 5 displayed search results, to keep the CA signal sufficiently reliable.

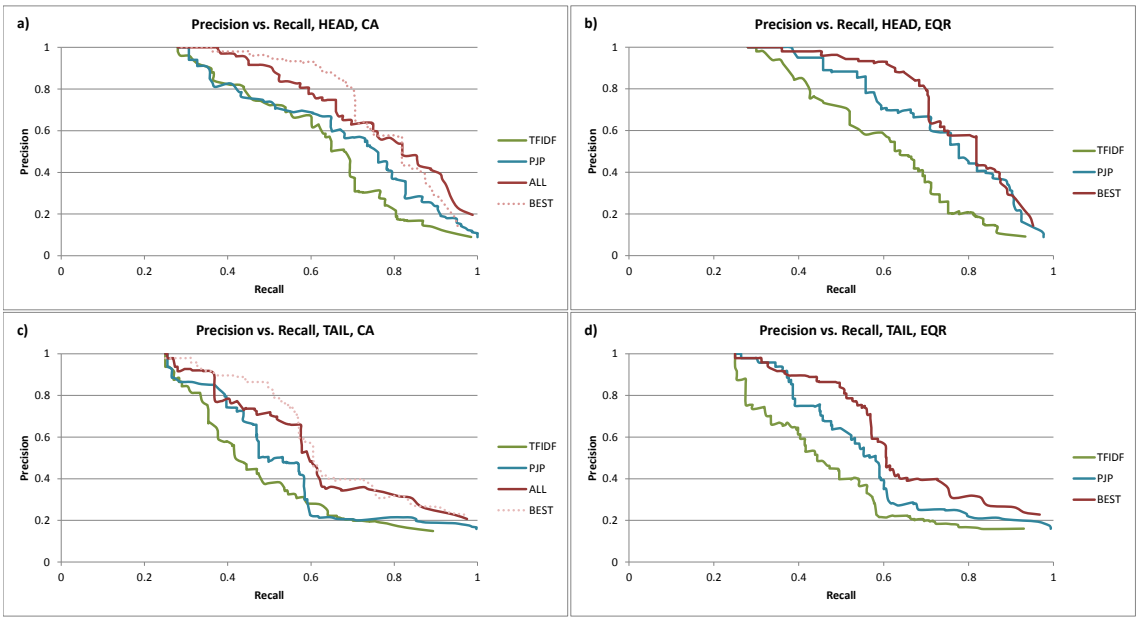

Figure 1: Comparison of precision vs. recall curves for: HEAD (a and b) and TAIL (c and d); and CA (a and c) and EQR (b and d) soft-labeling techniques. ALL outperforms PJP on all configurations and EQR soft-labeling outperforms CA on all configurations. BEST indicates our best configuration consisting of EQR soft labeling and ALL features.

it is extremely sparse in the tail; and (2) the head is represented dominantly in the training data. As (Paranjpe, 2009) points out, the strategy behind learning a salience model from a soft label is to learn from the cases where we have a good supervision signal and to generalize to the cases in the tail. Given this argument, our expectation was to see gains mostly in the tail for our proposed soft labeling function and feature set **EQR_ALL**. The positive gains on the HEAD were unexpected.

The results indicate that the choice of the soft labeling function is important. On TAIL, the trend is that EQR outperforms CA overall as a training signal. On HEAD, the soft labeling technique matters less; using our full feature set, both techniques yield similar performance except on F1 where EQR outweighs CA.

As discussed in Section 5.2, we consider both rank-sensitive applications and class-sensitive applications. As such we also test rank model configurations trained explicitly to optimize nDCG. Using our rank models, we observe par performance against the regression models on HEAD. On TAIL, the rank models outperform regression in the first position on nDCG and MAP.

Figure 1 shows the precision/recall characteristics on HEAD and TAIL, with CA and EQR soft labeling techniques combined with three different feature sets. The TFIDF features underperform compared to the other feature sets in all settings. The PJP feature set improves precision/recall in all cases against TFIDF, but in a more pronounced fashion when used with the EQR soft labels. The best precision/recall curves are obtained from **EQR_ALL**.

The dotted lines in (a) and (c) superimpose the best system curve (**EQR_ALL**) on the CA results, highlighting the observation that the best system produces precision/recall gains especially in the region where precision is greater than 0.7. At recall $\sim$ 0.6 the precision gain on `HEAD` is nearly 7.5 points, on tail it is nearly 10 points at $\sim$ 0.5 recall.

6.3 Contribution of Feature Families

6.3.1 Feature Weights

Examination of the feature weights in **EQR_ALL** reveals that the strongest salience cues are the position and the frequency of the entity in the document and anchor text. In the model, 174 features receive non-zero weights. The top five features are: the frequency of e in the anchor text, document and title, and the `df` of e and offset of e in the document. The next series of 37 features in order of feature weight is a mix of page classification, position, URL, structural and page segmentation features with no discernible prominence of any of these families. The binary features representing top level domains and page categories occur in the lower weight area of the feature list, with the exception of the binary feature indicating that the top level domain is Wikipedia - this feature ranks 11^{th} which is not surprising given the frequency and highly specific structure of this domain. Domains and page categories in the medium range of feature weights include: Amazon.com, News category, Arts category, Recreation/Outdoors category, Computers category etc.

6.3.2 Feature Ablation

Table 4 lists the results of a feature ablation study relative to our best model, **EQR_ALL**. Many features in **EQR_ALL** require either a sizeable web crawl or components that are typically part of a commercial search engine, such as page classifiers and page segmentation models. The first block in the table, **EQR_DOC**, addresses the question how well our salience model would perform without access to any features that rely on this information. This configuration would be relevant for a system that is trained outside of a search engine. Note, though, that the soft label is also unavailable without access to a web search log, so such a system would either have to rely on manually labeled data for training or design a new soft labeling function. On `HEAD`, the difference is minimal and not statistically significant, except for F1 where **EQR_ALL** outperforms **EQR_DOC**. In `TAIL`, however, the **EQR_ALL** achieves better results, with significant gains in nDCG@1 and nDCG@5.

The second block in Table 4 compares **EQR_ALL** with models trained with the individual feature families listed in Table 1. We omit the feature families that do not take into account the entity. These features are useful in conjunction with others, however they do not make sense to test separately since they will have identical output for all entities in a document. Few of the results are statistically significant, but we observe the following trends. On `HEAD`, the **Position** family fared best on all metrics except F1. On `TAIL`, **Position** and **Structural** families scored the highest.

	HEAD				
	nDCG@1	**nDCG@5**	**MAP@1**	**MAP@5**	**F1**
EQR_ALL	0.82	0.85	0.54	0.58	0.75
EQR_DOC	0.82	0.84	0.54	0.59	0.69†
EQR_Corpus	0.62†	0.74†	0.34†	0.44†	0.61†
EQR_Structural	0.72†	0.75†	0.42	0.50	0.66†
EQR_WebGraph	0.74	0.79	0.46	0.52	0.70
EQR_Position	0.78	0.82	0.48	0.53	0.69
EQR_PageSeg	0.74	0.80	0.46	0.50	0.64†
	TAIL				
	nDCG@1	**nDCG@5**	**MAP@1**	**MAP@5**	**F1**
EQR_ALL	0.73	0.77	0.48	0.58	0.64
EQR_DOC	0.68†	0.68†	0.42	0.50	0.58
EQR_Corpus	0.58†	0.63†	0.33†	0.44†	0.51†
EQR_Structural	0.75	0.65†	0.50	0.53	0.58
EQR_WebGraph	0.65†	0.59†	0.38†	0.46†	0.54†
EQR_Position	0.73	0.69	0.46	0.52	0.60
EQR_PageSeg	0.70	0.67†	0.44	0.48	0.54†

Table 4: Feature ablation experiments on `HEAD` and `TAIL` against our best performing model **EQR_ALL**. The first comparison block consists of only features that are not accessible to a typical commercial search engine, **EQR_DOC**. The second block corresponds to each of the feature families listed in Table 1 that involve the entity. † indicates statistical significance under the **EQR_ALL** configuration using Student's t-Test with p-value = 0.1.

6.4 Error Analysis

6.4.1 Soft Labeling Errors

When analyzing CA and EQR and their scores on various entities, the following observations emerge: CA's high scores (> 0.8) and zero scores conform well to human judgment, while it shows more errors in the range of non-zero scores to mid-level scores. For example, the 1994 song "Dummy Crusher" by Kerbdog receives a CA score of only 0.05 on the Wikipedia page with the same title, while its EQR score is 1. "Dummy Crusher" happens to also be a popular online game, which likely means that many users who issue the query are looking for the game rather than the song, hence they will skip the Wikipedia link in favor of the "Play Dummy Crusher" game URL. EQR, on the other hand, is based on the information that all of the users who land on the Wikipedia page after a query have issued a "Dummy Crusher" query. While this set of users may be very small compared to the ones who are looking for the game, EQR is not influenced by popularity.

A closer look at the cases where CA assigns a low score to an entity that is actually salient on the page leads to the hypothesis that two main reasons may contribute to this error. First, popular entities such as celebrities are likely to trigger video and image search results in the SERP, which may cause the user to skip the Wikipedia page in favor of the more visually attractive content. Second, news and entertainment articles about celebrities may

be skipped in favor of image galleries or more comprehensive web pages about the celebrity, their currently popular TV show, etc. In general, the "skipping" behavior exploited by CA is influenced by many factors of search result presentation and popularity that are independent of the aboutness of the displayed pages, as hypothesized in Section 4.2.1.

For EQR, we again observe that the high scores correspond nearly without error to the human assessment. The mid- and low range of scores, however, show better reliability than in CA. For example, a celebrity who is the subject of an entertainment news article is correctly scored highly by EQR. One notable source of low score errors for EQR is spelling or naming variants. "Margarita Island" gets a high salience score on the Wikipedia page for "Isla Margarita", while "Isla Margarita" itself gets a low score, presumably because the Spanish original term is not used as often in English queries as the English variant.

To summarize, while EQR proves to be the more reliable soft labeling method, both EQR and CA have shortcomings: CA is unreliable when external factors influence the click/skip behavior, such as position bias and popularity. EQR, on the other hand, suffers from spelling and naming variations. Note, however, that the latter cause for error is more easily remedied (for example by conflating queries/entities that have a strong lexical similarity).

6.4.2 Prediction Errors

We turn now to the prediction errors in **EQR_ALL** versus **CA_PJP**. We find that in those cases where the two models make significantly different predictions, **EQR_ALL** often has the higher entity salience score. Our features tend to lead to improved performance in cases where a salient entity is hidden in a less conspicuous position, e.g., when it is not present in the title or in the first paragraph. In these cases, **CA_PJP** often underestimates salience. A common problematic case for both models is the presence of salient entities in banners on the web page. While these entities are typically not salient (e.g. the entity "BBC News" on the banner of a news story), there are exceptions when the page is actually the home page for the entity. While our current features are not sufficient to make the proper distinctions, more sophisticated features could potentially solve this problem. For example, the structure of the text on the page (article-type, multi-paragraph, multi-sentence text versus menu/navigation-type content) might be a helpful cue in these cases. Not unexpectedly, both models also struggle when a web page contains multiple lexical variants of an entity's name (e.g. the baseball star *Troy Tulowitzki* and his nickname *Tulo*), indicating that adding coreference resolution for entities (and pronouns) to our system would help.

7. Conclusion

Mining meaningful connections between entities and content has proven difficult because of the large percentage of entities that are not salient or central to the documents in which they occur. This paper formalizes and addresses the task of ranking and classifying entity-URL pairs according to the salience of the entity in the document. We propose a system with the following properties: the system is cost-effective to build and improves upon the state of the art. To eliminate the need for expensive manual annotations, we propose weakly-supervised learned models combined with a novel method for automatically labeling large quantities of training data by leveraging usage behaviors found in web search logs. This, along with

an extensive feature set leads to significant improvements over the current state of the art on both head and tail distributions of the web. As no public data exists to date to evaluate this task, we design and release to the research community a gold standard data set with salience annotations, representing the head and tail distributions of pages on the web.

References

Agichtein, E., Brill, E., & Dumais, S. T. (2006). Improving web search ranking by incorporating user behavior information. In *Proceedings of SIGIR*.

Baeza-Yates, R. (2004). Web usage mining in search engines. *Web mining: applications and techniques*, 307–321.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *the Journal of machine Learning research*, *3*, 993–1022.

Bruza, P. D., Song, D. W., & Wong, K. F. (2000). Aboutness from a commonsense perspective. *Journal of the American Society for Information Science*, *51*, 1090–1105.

Cai, D., Yu, S., Wen, J., & Ma, W. (2003). Extracting content structure for web pages based on visual representation. *Web Technologies and Applications*, 406–417.

Collins, M. (2002). Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *Proceedings of EMNLP*.

Cucerzan, S. (2007). Large-scale named entity disambiguation based on wikipedia data. In *Procceedings of EMNLP-CoNLL*.

Dalvi, N. N., Kumar, R., Pang, B., Ramakrishnan, R., Tomkins, A., Bohannon, P., Keerthi, S., & Merugu, S. (2009). A web of concepts. In *Proceedings of PODS*.

Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL*.

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

Hjørland, B. (2001). Towards a theory of aboutness, subject, topicality, theme, domain, field, content... and relevance. *Journal of the American Society for Information Science and Technology*, *52*(9), 774–778.

Holland, S., Ester, M., & Kießling, W. (2003). Preference mining: A novel approach on mining user preferences for personalized applications. *Knowledge Discovery in Databases: PKDD 2003*, 204–216.

Hovy, E., & Lin, C. Y. (1998). Automated text summarization and the summarist system. In *Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998*.

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of EMNLP*.

Hutchins, W. (1977). On the problem of 'aboutness' in document analysis. *Journal of Informatics*, *1*(1), 17–35.

Irmak, U., Brzeski, V. V., & Kraft, R. (2009). Contextual ranking of keywords using click data. In *Proceedings of ICDE*.

Joachims, T., Granka, L., Pang, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of SIGIR*.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of SIGKDD*.

Komachi, M., & Suzuki, H. (2008). Minimally supervised learning of semantic knowledge from query logs. In *Proceedings of IJCNLP*.

Kupiec, J., Pedersen, J. O., & Chen, F. (1995). A trainable document summarizer. In *Proceedings of SIGIR*.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Marcu, D. (1997). From discourse structures to text summaries. In *Proceedings of ACL*.

Muralidharan, A., Gyongyi, Z., & Chi, E. H. (2012). Social annotations in web search. In *Proceedings of SIGCHI*.

Pantel, P., Gamon, M., Alonso, O., & Haas, K. (2012). Social annotations: Utility and prediction modeling. In *Proceedings of SIGIR*.

Paranjpe, D. (2009). Learning document aboutness from implicit user feedback and document structure. In *Proceedings of CIKM*.

Paşca, M., & Durme, B. V. (2007). What you seek is what you get: Extraction of class attributes from query logs. In *Proceedings of IJCAI*.

Putnam, H. (1958). Formalization of the concept 'About'. *Philosophy of Science*, *25*(2), 125–130.

Radlinski, F., & Joachims, T. (2005). Query Chains: Learning to rank from implicit feedback. In *Proceedings of SIGKDD*.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513–523.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620.

Salton, G., Allan, J., & Buckley, C. (1993). Approaches to passage retrieval in full text information systems. In *Proceedings of SIGIR*.

Song, R., Liu, H., Wen, J., & Ma, W. (2004). Learning block importance models for web pages. In *Proceedings of WWW*.

Xu, G., Yang, S., & Li, H. (2009). Named entity mining from click-through data using weakly supervised latent dirichlet allocation. In *Proceedings of SIGKDD*.

Yih, W., Goodman, J., & Carvalho, V. (2006). Finding advertising keywords on web pages. In *Proceedings of WWW*.

Zhu, X. (2005). Semi-Supervised Learning Literature Survey. Tech. rep. 1530, Computer Sciences, University of Wisconsin-Madison.