Enhancing Web Search with Entity Intent

Na Dai Xiaoguang Qi Brian D. Davison
Dept. of Computer Science & Engineering, Lehigh University
Bethlehem, PA 18015 USA
{nad207,xiq204,davison}@cse.lehigh.edu

ABSTRACT

Web entities, such as documents and hyperlinks, are created for different purposes, or intents. Existing intent-based retrieval methods largely focus on information seekers' intent expressed by queries, ignoring the other side of the problem: web content creators' intent. We argue that understanding why the content was created is also important. In this work, we propose to classify such intents into two broad categories: "navigational" and "informational". Then we incorporate such intents into traditional retrieval models, and show their effect on ranking performance.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithms, Performance

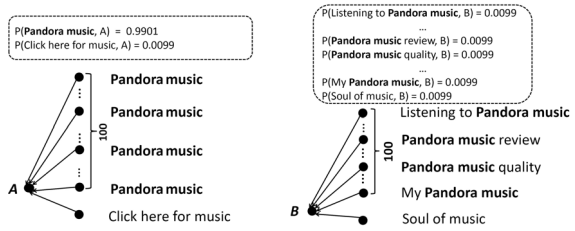
Keywords: web entity intent, query intent, retrieval model

1 Introduction

Understanding users' intent is important for search engines to better serve users' information needs. Broder proposed a taxonomy for such intents, in which queries are classified into three categories: navigational, informational, and transactional [3]. While previous research showed that query intent classification can improve ranking performance, they ignored the connection to the intents of information providers when creating web content. As illustrated in Figure 1, we argue that the intent behind every hyperlink can influence its importance to the target page, and thus affect ranking effectiveness. In this work, we propose to model the intent for which web pages and hyperlinks (links for short) are created, and incorporate such intent into ranking methods to show its usefulness.

People create links for various reasons. In this brief work, we categorize links into two types: (1) *navigational links*: links that are created to describe the target page's identity; and (2) *informational links*: links created to describe the target. For example, a link pointing to "http://www.facebook.com/" with the anchor text "Facebook" is considered a navigational link since the anchor text is the proper name of that particular web site. A link with the same anchor text pointing to the Wikipedia page describing Facebook Inc. is an informational link. These two classes are not exclusive. We consider every link to be a soft combination of both. Similarly, we classify web pages into "navigational intent pages" (pages that mostly attract navigational links) and "informational intent pages" (pages that mostly attract informational links). Our work is conducted in two steps: (1) classify links into the two intent classes; (2) use the link classification result to generate better rankings. Our key contribution is to show the effect on ranking when incorporating the intents of the creation of web entities (links and pages) into retrieval models.

Copyright is held by the author/owner(s).
WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0632-4/11/03.



(a) Page A is the home page of *pandora music*.

(b) Page B is a blog discussing music on Site *pandora music*.

Figure 1: Two pages have 101 inlinks, and 100 anchortexts contain *pandora music*. As a result, the contributions of anchor terms *pandora music* to A and B are indistinguishable without considering the page content and anchor text length. However, we claim that Page A should be emphasized when answering the navigational query "*pandora music*", and the links with *pandora music* pointing to A are more navigational.

2 Link Intent Classification

To incorporate link intent into ranking, the first step is to train a classification model to identify link intents. The approach is summarized below. Our extended report [5] has details.

Labels: Each anchor-URL pair in the training set is labeled by human as either "navigational" (positive class) or "informational" (negative class).

Features: The features for link intent classification include numerical features that quantify URL length and depth, anchor text length, statistics based on the Part-Of-Speech (POS) tags of anchor terms, e.g., fractions of nouns, verbs, etc.; as well as anchor entropy information, such as features of the entropy based on anchor term distribution, defined as $i(a_t) = \frac{H(D_{a_t}|a_t)}{\log|D_{a_t}|}$, where $H(D_{a_t}|a_t) = -\sum_{d\in D_{a_t}} P(d|a_t)\log P(d|a_t)$ represents the conditional entropy of the distribution on URL collection (D_{a_t}) associated with anchor term a_t.

Model: A SVM classifier based on LIBSVM [4] is trained using the labeled training set. Given an unseen pair, the model generates the probability that its associated link has "navigational intent". So each link l corresponds to a binomial intent distribution $\overrightarrow{\mathbf{I}_l}$, with each item representing the probability that l has specific intent. Page intent distribution is the centroid of all its incoming link intent distributions, defined as $\overrightarrow{\mathbf{I}_d} = \frac{1}{|Inlink(d)|}\sum_{l\in Inlink(d)}\overrightarrow{\mathbf{I}_l}$.

3 Using Entity Intents in Retrieval Models

Given the web entity intent distribution, we next incorporate them into two representative retrieval models, i.e., language model (LM) and BM25F [6]. While our intent taxonomy only includes two

Table 1: Link intent classification performance (577 out of the 1000 selected link instances are labeled as "navigational" or "informational", with a "navigational/informational" ratio of 0.2516).

Class	Precision	Recall	F_1-measure	Accuracy
Navigational	0.7680	0.8275	0.7966	0.9150
Informational	0.9557	0.9370	0.9463	0.9150

Table 2: Ad hoc task performance on TREC Web Track 2009. Performance with significant improvement (p-value$<$0.05) over baseline is marked as †. Improvement with p-value$<$0.01 over baseline is marked as ‡.

Incorporating Intent into LM					
Sim	statMAP	P@1	P@3	P@10	P@20
Baseline	0.1758	0.2040	0.2244	0.3040	0.3447
1-LD/2	0.1767	**0.2244**	0.2312	**0.3381**†	**0.3679**†
Cosine	**0.1771**	0.2040	**0.2380**	0.3126	0.3376
Incorporating Intent into BM25F					
Sim	statMAP	P@1	P@3	P@10	P@20
Baseline	0.1752	0.2857	0.3061	0.3777	0.3892
1-LD/2	0.1738	0.3469‡	**0.3265**†	0.3916	0.3928
Cosine	**0.1756**	**0.3673**‡	0.3061	**0.4249**‡	**0.4019**†

classes, the idea can be generalized to retrieval models based on other taxonomies. We emphasize two aspects: (1) combining lexical similarity with intent similarity between queries and web pages; and (2) incorporating intent similarity between incoming links and target pages into target page representation.

Incorporating intent into LM: The probability that query q generates web page d is estimated by $p(d|q) = p(q|d)p(d)$, where $p(d)$ is constant for all pages since no page is more relevant than others for all queries. Each page d is represented by $d =< C_d, \overrightarrow{\mathbf{I}_d} >$, where C_d is d's content. Assuming page content and intent are independent, $p(d|q)$ is given by: $p(d|q) = p(C_d, \overrightarrow{\mathbf{I}_d}|q) \propto p(q|C_d)p(\overrightarrow{\mathbf{I}_d}|q)$ where $p(q|C_d)$ is the query likelihood, a well-studied factor in previous work. $p(\overrightarrow{\mathbf{I}_d}|q)$ is the probability that q generates a page with the same intent as d. We make the assumption that $p(\overrightarrow{\mathbf{I}_d}|q)$ is proportional to $sim(\overrightarrow{\mathbf{I}_q}, \overrightarrow{\mathbf{I}_d})$, where $\overrightarrow{\mathbf{I}_q}$ is the query intent estimated by accumulating the intents of pseudo-feedback documents, i.e., the top k search results ($k = 100$) generated by $p(q|C_d)$ in this work.

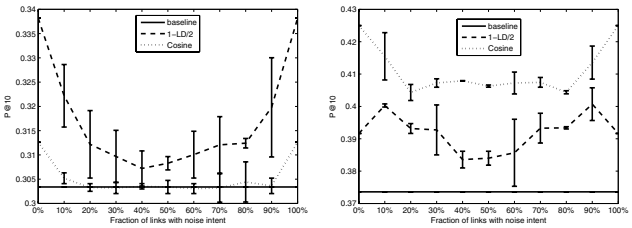
Incorporating intent into BM25F: BM25F linearly combines term frequencies in multiple page fields. The frequency of term t in page d's anchor field can be estimated by: $w_{anchor}(t, d) = \sum_{a \in A(d)} w(a, d)tf(t, a)$ where $w(a, d)$ is the importance of anchor a to page d. When only considering anchor fields, d's term frequency is given by: $w(t, d) = (1 - \alpha)w_{body}(t, d) + \alpha w_{anchor}(t, d)$. Assuming the incoming links sharing similar intent distribution with target pages are more important, we incorporate an intent similarity measure into anchor importance estimation by $w'(l, d) = sim(\overrightarrow{\mathbf{I}_l}, \overrightarrow{\mathbf{I}_d})w(l, d)$. Note that page length can be calculated in a similar way.

Estimating $sim(\overrightarrow{\mathbf{I}_i}, \overrightarrow{\mathbf{I}_j})$: We estimate $sim(\overrightarrow{\mathbf{I}_i}, \overrightarrow{\mathbf{I}_j})$ by using either (1) Cosine similarity between $\overrightarrow{\mathbf{I}_i}$ and $\overrightarrow{\mathbf{I}_j}$; or (2) $(1 - LD/2)$ where LD is the L_1 distance between $\overrightarrow{\mathbf{I}_i}$ and $\overrightarrow{\mathbf{I}_j}$.

4 Evaluation and Conclusion

Our goal is to improve search quality by incorporating web entity intent into retrieval models. The experiments are conducted on ClueWeb09 (Category B) data set, with 49.8M web pages and 940M links. We use the 50 queries in the Ad hoc task of the TREC 2009 Web track for evaluation.

To generate anchor-target pairs for link intent classification, we first split 50 queries into 5 folds sequentially by query IDs. For each query, we collect its top-1000 documents by query likelihood. We randomly select 200 in-links pointing to pseudo-feedback documents in each fold as link intent instances (no overlap between folds). Each instance is labeled by at least one worker of Amazon Mechanical Turk [2], in selection among "navigational", "informational", "none of them" or "both of them". We use only the instances labeled as "navigational" or "informational" for training. Table 1 shows the classification performance based on 5-fold CV.



(a) LM (b) BM25F

Figure 2: Performance on P@10 varying the fraction of noise in link intent.

For ranking evaluation, we compare the retrieval models after incorporating link intents with the baselines, i.e., LM and BM25F. statMAP [1] and Precision at truncation level k (P@k) are our main metrics. For language model experiments, the query likelihood is estimated by $p(q|d) = \prod_{w_i \in q} p(w_i|d)$. Dirichlet smoothing is used for $p(w_i|d)$ estimation, with smoothing parameter μ being 2500. For BM25F experiments, $w(a, d)$ is simply the number of times that anchor a points to document d. All methods are compared when the trade-off parameter α equals 0.5.

The ranking performance comparison is shown in Table 2. After incorporating web entity intents, both LM and BM25F outperform baselines on most metrics. It suggests that introducing a bias based on web entity intents helps improve retrieval quality. We attribute such improvements to a better connection among the intents of providers, recommenders and seekers of information. On BM25F, our approach can better improve the quality of top results, which is often of particular value in web search.

To better understand the effect of link intent on ranking, we intentionally introduce noise into link intent classification result by reversing $\mathbf{I}_l$ of a fraction of randomly selected links and conducted 30 runs at each noise level. The average and deviation of ranking performance on P@10 is shown in Figure 2. We observe that: (1) ranking performance decreases with the increase of noise (note that the trends are approximately symmetric with respect to the noise level at 50% since the retrieval models equally treat two types of link intents); and (2) the retrieval models using *Cosine* similarity measure is more tolerant to noise.

In summary, we showed that incorporating web entity intents into retrieval models can improve retrieval quality. The improvements are sensitive to the accuracy of link intent classification.

Acknowledgments

This work was supported in part by grants from the National Science Foundation under award IIS-0803605 and IIS-0545875, and an equipment grant from Sun Microsystems.

5 References

[1] J. Allan, B. Carterette, B. Dachev, J. A. Aslam, V. Pavlu, and E. Kanoulas. Million query track 2007 overview. In *TREC*, 2007.

[2] Amazon Inc. Amazon mechanical turk. `http://www.mturk.com/`, 2010.

[3] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, Fall 2002.

[4] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[5] N. Dai, X. Qi, and B. D. Davison. Bridging link and query intent to enhance web search. Tech. Rep. LU-CSE-10-006, Dept. of Computer Science and Engr., Lehigh Univ., 2010.

[6] S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In *ACM CIKM 2004*, pages 42–49, 2004.