Social Annotations: Utility and Prediction Modeling

Patrick Pantel, Michael Gamon
Microsoft Research
One Microsoft Way
Redmond, WA, USA
{ppantel,mgamon}@microsoft.com

Omar Alonso, Kevin Haas
Microsoft Corp
1065 La Avenida
Mountain View, CA, USA
{omalonso,kevinhaa}@microsoft.com

ABSTRACT

Social features are increasingly integrated within the search results page of the main commercial search engines. There is, however, little understanding of the utility of social features in traditional search. In this paper, we study utility in the context of social annotations, which are markings indicating that a person in the social network of the user has liked or shared a result document. We introduce a taxonomy of social relevance aspects that influence the utility of social annotations in search, spanning query classes, the social network, and content relevance. We present the results of a user study quantifying the utility of social annotations and the interplay between social relevance aspects. Through the user study we gain insights on conditions under which social annotations are most useful to a user. Finally, we present machine learned models for predicting the utility of a social annotation using the user study judgments as an optimization criterion. We model the learning task with features drawn from web usage logs, and show empirical evidence over real-world head and tail queries that the problem is learnable and that in many cases we can predict the utility of a social annotation.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Design, Experimentation, Measurement

Keywords

Social relevance, social aspects, web search

1. INTRODUCTION

Social offerings are becoming table stakes for the major search engines. Querying for your local pizza restaurant on Bing or Google may yield reviews from your friends about the restaurant, or about nearby restaurants liked by your friends. A query for "*maui hotels*" may yield a result indicating that your colleagues like the *Royal Lahaina Resort* on Facebook and a query for "*luau history*" may yield a social annotation on the result "Hawaii Luau History" stating that your friend has liked or shared the document.

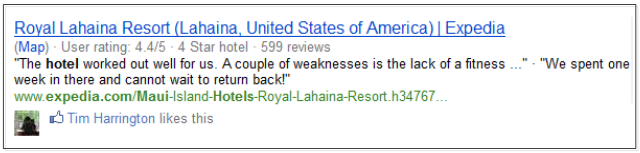

Figure 1: Example social annotation for the query "*maui hotels*".

The user can benefit from such social experiences in various ways, including: (a) discovery of socially vetted recommendations; (b) personalized search results; (c) connecting to the lives of their friends; (d) result diversity; and (e) emotionally connecting with an otherwise static and impersonal search engine. There is, however, very little understanding whether these social features are useful or detrimental to the whole-page user experience, or, for that matter, how to even measure the utility of social features.

Consider a **social annotation** feature, such as the one depicted in Figure 1, where some results on the search results page (SERP) are enriched with markings indicating some of your friends that have previously `liked` or `shared` that result[1]. Are such endorsements from dearest friends more relevant to the user than from acquaintances or coworkers? Are expert opinions or those from friends who live in the vicinity of the restaurant more valuable? Do annotations on irrelevant results amplify their negative perception?

Studying such aspects and their effect on social relevance form the basis of this paper. We begin by enumerating a taxonomy of social relevance aspects, i.e. cues or criteria that influence the perceived utility of social annotations. We consider aspects related to the user query, the social connection, and the relevance of the related content (e.g., a document returned by the search engine). We then define measures of social annotation utility and present the results of a large controlled user study of the interplay between each of these aspects on head and tail queries drawn from several months of real-world queries issued to a commercial search

[1]Social networks have different terms to indicate explicit positive user interest of a document, such as *like*, *+1*, and *tag*. Herein we use the term `like` generically to indicate any of these, and `dislike` to indicate negative interest.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.

engine. We show evidence that social annotations add perceived utility to users in varying degrees according to the social relevance aspects. Finally, we turn our attention to the task of automatically predicting the utility of a social annotation. We model the prediction task using a multiple additive regression tree model over features available to a standard search engine. We show empirical evidence that the task is learnable and that we can automatically predict the utility of a social annotation. The major contributions of our research are:

- We introduce a taxonomy of social relevance aspects, drawn from the query, social connection and content, which influence the utility of social annotations on the search results page.
- We conduct a user study to quantify the influence and interplay of the social relevance aspects, over real-world social annotation impressions drawn from a commercial search engine.
- We propose a machine learned discriminative model for predicting the expected added value of a social annotation in an online scenario.
- We empirically show that our model can accurately predict the relevance of a social annotation.

2. RELATED WORK

Most relevant to our work is that of Muralidharan et al. [18] who show through user studies and eyetracking analysis that the presentation of the social annotation, such as the size of the profile thumbnail, greatly impacts user engagement of the annotation. Through anecdotal feedback, participants from their studies conjectured that annotations would be useful in certain social and subjective topics (e.g., restaurant and shopping queries) or when presented by friends believed to be topical authorities (e.g., a fitness trainer annotating a fitness web document) or when in a close relationship with the searcher; but for other situations the participants believed annotations were not helpful, such as when there is no label explaining why an annotation was present. In this paper, we build upon their work by exploring a taxonomy of influential social relevance aspects, including query class, content relevance, and social network aspects; and quantifying their utility and interplay.

2.1 Social Features in Web Search

Evans and Chi [9] performed a detailed user study about the role that social interactions play in collaborative search tasks. Outside of collaborative search, social signals have primarily been explored as implicit ranking features hidden deep inside the ranking functions [20, 26], for example by examining how users would benefit from personalized search results considering implicit behavior similarity attributes such as from click-based measures [22, 23, 1]. Bao et al. [2] further argue that the quality of a web page can be improved by the amount of del.icio.us annotations and Carmel et al. [6] show that personalized search improves the quality of intranet search results. For traditional web search, Heymann et al. [13] predicted that while social bookmarks can provide ancillary data not present in the web page, the majority of tags are also present in the document, or inlinks/outlinks, and therefore would have limited use as ranking features.

Recently, social features are appearing as explicit user-facing features such as in: (1) **annotations**, where interest by the searcher's social network is visibly marked on an existing search result (v.s. Figure 1); (2) **injected results**, where social data, such as tweets or status posts, are presented in a manner similar to and within the existing search results, but sourced outside of the web corpus; and (3) **independent results**, where the social data is presented in a manner not to be mistaken for one of the web search results, such as in a web answer or direct display. This paper specifically focuses on measuring utility in the context of social annotations and leaves the analysis of other visualizations to future work.

2.2 Measuring the Utility of a Search Result

Relevance is a multidimensional dynamic concept and there is a wide range of factors that influence a user's perception of relevance. Via an extensive user study, Barry and Schamber proposed a set of categories capturing users' relevance criteria or *cues* (e.g., accuracy, specificity, expertise, presentation, etc.) in the context of an information seeking task [3]. More recently, Borlund provided a framework for examining relevance in the context of information retrieval evaluation [4]. The relevance and usefulness of results returned by web search engines are typically evaluated using variants of nDCG [16], expected reciprocal rank [7], mean average precision, relative information gain, and a variety of click/feedback metrics [25, 15], such as (1) clicks on lower-ranked documents indicate that higher-ranked documents are less relevant for the query; and (2) clicks to documents which are quickly abandoned by the user for other search results are deemed less relevant for the query. However, use of these traditional metrics can present challenges when personalizing and annotating web search results, as higher-ranked search results may be passed up for lower-ranked search results with social annotations. As shown in [8] and [12], annotations and other modifications to captions can alter the success rate for users independent of where the document was ranked in the result set. Fidel and Crandall [10] show factors beyond the document that affect the perception of relevance, including recency, detail, and genre; but they do not discuss social factors. This paper extends their work and also proposes metrics by which the utility of a social annotation feature can be measured.

2.3 Predicting Relevance

At the core of a search engine is the ability to learn to rank candidate documents according to their relevance to a user query. As such there is a plethora of work on modeling, feature extraction, selection, and model adaptation, see Liu [17] for a comprehensive survey and Burges et al. [5] for a description of the system that won the recent Yahoo! Learning to Rank Challenge. In our work, we introduce a complementary task, that of learning to predict the relevance of a social annotation. As such, we make use of a state-of-the-art learning algorithm [27] and predict social relevance based on runtime features from web usage logs and query classifiers commonly used in web search ranking, as well as social relevance cues defined in this paper.

3. ASPECTS OF SOCIAL RELEVANCE

Consider a search results page consisting of ranked content (e.g., documents, videos, images) in response to a user

query. Formally, we define a **social annotation** as a tuple, $\{q, u, c, v\}$, consisting of a query q, content u, a social network connection c and the connection's interest valence v in the content (e.g., `like`, `dislike` or `share`). For example, Figure 1 illustrates such a social annotation impression where q is "*maui hotels*", u is a relevant Expedia hotel page, c is "*Tim Harrington*" and v is `like`.

In this section we propose a taxonomy of aspects that influence the utility of a social annotation, spanning the query, the social connection, and the content.

3.1 Query Aspects (QA)

3.1.1 Query Intent (QA-INT)

We divide queries into two sets based on whether they are **navigational (nav)** or **non-navigational (nnv)**. We expect that social annotations will be less valuable when the user is simply looking for the url of a web page she wants to reach. Other possible intents that are not explicitly studied in this paper include informational and transactional intents, which we grouped within our `nnv` intent, and which partially overlap with our `QA-CLS` aspects described next.

3.1.2 Query Class (QA-CLS)

The utility of a social annotation may be influenced by the class of the issued query. For example, although one might find value in knowing the interest of a social connection when querying for a movie, song, or book, we may expect only certain connections such as experts to be of interest for a health query. Similarly for local queries, interest from connections in the vicinity of the target location are likely more valuable than distant connections.

We focus our analysis on the following query classes covering the majority of social annotations found on a commercial search engine:

- **Commerce (com)**: Product-related queries seeking specifications, prices, comparisons, transactions, and reviews.
- **Health (hea)**: Queries on health-related topics such as symptoms, procedures, and treatments.
- **Movies (mov)**: Movie title queries.
- **Music (mus)**: Queries for musical artists, bands, and song lyrics.
- **Restaurant (res)**: Local queries related to restaurants, cafes, and drinking establishments.

3.2 Social Connection Aspects (SA)

The social network connection is at the heart of the social annotation and it clearly influences its utility. Consider the query "*korean restaurant*" and a set of resulting links to nearby korean restaurants. We expect that a connection who is an expert on korean cuisine may increase the utility of an annotation, and that a connection living near the localized neighborhood is more important than one far away. A dear friend's interest in the restaurant may also hold more weight than a more distant work colleague. And finally, the interest valence, whether positive, negative or neutral might affect relevance. Below we summarize each of these aspects and their value ranges.

3.2.1 Circle (SA-CIR)

The circle of a connection refers to the relation of the connection to the searcher. Intuitively, a co-worker's interest in an article related to the workplace might hold more value than a family member or friend's interest. We consider the following circles: **work colleague (wkc)**, **family member (fam)**, and **friend (frn)**. Other interesting circles out of scope for this paper include school friends, college friends, church friends, and sports club friends.

3.2.2 Affinity (SA-AFF)

The affinity between a searcher and a connection refers to their degree of closeness. Although affinity has a continuous range, in this paper we consider two affinities: **close (cls)** and **distant (dst)**. As conjectured in [18] we generally expect the closeness of a connection to greatly influence the perceived utility of a social annotation.

3.2.3 Expertise (SA-EXP)

Whether a connection is an **expert (exp)** or **non-expert (nex)** on the search topic may influence the value of a social annotation especially when issuing informational or transactional queries. Other possible values not considered in this paper include hobbyist and enthusiast.

3.2.4 Geographical Distance (SA-GEO)

For local queries, we expect that a connection living **near (nea)** the target location will add more value than if living **far (far)**. For non-local queries, the value of this aspect is **not applicable (n/a)**. One might also consider the geographic distance between the searcher and the connection, but we leave this aspect for consideration in future work.

3.2.5 Interest Valence (SA-INT)

Social annotations require both a social network connection and the interest valence of that connection with respect to the annotated search result. Most experiences today classify the valence as either a **like (lik)** or a **share (shr)**, i.e. that the user has shared the document with someone. In this paper, we also consider a third valence, **dislike (dis)**.

3.2.6 Out of Scope Aspects

Other aspects may influence the value of a social annotation, such as the connection's gender, age, and general interests. We leave these for further study in future work.

3.3 Content Aspects (CA)

Finally, the social annotation is influenced by the relevance of the document to the intent of the user query. We consider graded relevance according to the following scale, similar to that proposed by Järvelin and Kekäläinen [16]:

- **Perfect (per)**: The page is the definitive or official page that strongly satisfies the most likely intent.
- **Excellent (exc)**: The page satisfies a very likely or most likely intent.
- **Good (goo)**: The page moderately satisfies a very likely or most likely intent.
- **Fair (fai)** The page weakly satisfies a very likely or most likely intent.
- **Bad (bad)**: The page does not satisfies the intent.
- **Detrimental (det)**: The page contains content that is inappropriate for a general audience.

4. SOCIAL RELEVANCE

This section presents the results of a user study quantifying the utility of a social annotation. We analyze the interplay between the social relevance aspects presented in Section 3 and identify situations where a social annotation is more relevant than others. Our approach is to sample social annotation impressions on a commercial search engine and to have human annotators judge the utility of each impression. In the following sections, we describe our process for sampling social annotation impressions (i.e., $\{q, u, c, v\}$-tuples) and the guidelines for judging their relevance. Then, in Sections 4.4 and 4.5, we present our analysis.

4.1 Query-URL Sampling

We define $\mathcal{U}$ as the universe of all social annotation impressions observed on a commercial search engine during three weeks spanning three months of US English web usage logs: 10/7/2011-10/14/2011, 11/4/2011-11/11/2011, and 12/2/2011-12/9/2011. We admitted only queries that were classified by the search engine to be in our domains of interest, namely commerce, health, movies, music, or restaurants (see Section 3.1.2). We applied a low classification threshold to admit a larger number of queries since we later manually annotate the query class of all queries in our test set. We further rejected any query suspected of being bot-generated.

We define the head of $\mathcal{U}$ as all tuples where q is in the top-20% of the most frequent queries, and the tail as all tuples where q is in the bottom 30%. A query-frequency weighted sample of queries from both sets yield our test queries, referred to as `HEAD` and `TAIL` consisting of 2388 and 1375 queries respectively.

For each query in `HEAD` and `TAIL`, we randomly selected one url from $\mathcal{U}$ using an impression-weighted sampling (i.e., a url with many social annotation impressions in the usage logs for the query is more likely to be chosen than a url with fewer social annotation impressions for the query).

We used a crowdsourcing tool to have each query in `HEAD` and `TAIL` manually classified according to the query classes defined in Section 3.1.2 as well as an `other` class. We obtained three judgments per query (7526 judgments) from a total of 32 independent annotators and kept the majority vote. The inter-annotator agreement as measured by Fleiss' κ was 0.495 (moderate agreement).

We further manually judged the relevance of each url to the associated query in `HEAD` and `TAIL`. The relevance categories and guidelines are those listed in Section 3.3. As this task is known to be more difficult than query classification, we employed seven professional independent annotators with experience in search engine relevance testing. The observed inter-annotator agreement as measured by Fleiss' κ was 0.176 (slight agreement).

4.2 Social Annotation Sampling

In order not to bias the social annotations to the specific social networks of our human annotators, we simulated a social network for our judges. We used this network to create the connections c and interest valences v for our random query-url pairs in `HEAD` and `TAIL`.

Figure 2 illustrates the virtual social network. It consists of twelve connections spanning the social circles and affinities defined in Section 3.3. We assume in this network that `family members` always have a `close` affinity whereas `work colleagues` have a `distant` affinity. `Friends` can have affinity either `close` or `distant`.

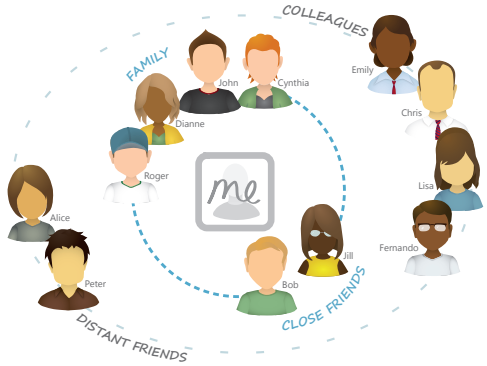

Figure 2: Simulated social network.

For each query-url pair in `HEAD` and `TAIL` we assigned a social annotation as follows. First, we randomly sampled a circle (either `work colleague`, `family`, or `friend`). Then, we randomly selected an individual from that circle, which also determines the affinity of the connection (either `close` or `distant`). Next, we randomly picked whether the connection was an `expert` or `non-expert` with respect to the document and we randomly chose the interest valence (either `like`, `dislike` or `share`). Finally, if the query was annotated as a local query, we randomly determined if the connection lived `near` or `far` from the intended target location. Otherwise, we set the geo-distance aspect to `n/a`.

Deploying our user study over the actual social networks of the participants is preferable, however several problems arise that make this infeasible, most notably privacy concerns. Also, since our study requires significant training and expertise, we found it necessary to hire professional annotators, thus limiting the total number of independent annotators. Using their personal networks would not only cause privacy concerns, but it would also bias towards a non-representative set of search users. Simulating a social network as we do carries its own risks. Firstly, we expect people's personal social networks to vary in terms of attribute value distributions (for example, some people have only work colleagues in their network while others have mostly friends and family) as well as diversity distribution. Although in our setup we assumed uniform priors for all attributes, if given the true priors it is trivial to reweight the findings. Secondly, the degree of an individual in their social network (i.e., average number of network connections) is significantly larger than the twelve in our virtual network, and also varies significantly[2]. To balance the cognitive load on our judges, the reliability of judgments, and reducing the risks of simulating a social network, we chose to keep the connection degree small enough whilst ensuring diversity in age, gender, and ethnicity. Finally, by explicitly drawing the judges' attention to social aspects, there is a risk of overemphasizing their importance and thus influencing the judgments, though we expect this influence to be minimal.

4.3 Annotation Task

The judges were presented with the following scenario:

[2]Hill and Dunbar [14] estimate the average connection degree at 153 and Ugander et al. [24] measured the median degree of Facebook users in May 2011 at 99.

	HEAD			TAIL		
	Oct'11	Nov'11	Dec'11	Oct'11	Nov'11	Dec'11
Test Cases	770	823	795	423	486	466
Judgments	1540	1646	1590	846	972	932
sig-util	402	634	428	273	363	225
some-util	734	610	785	397	341	449
no-util	364	377	315	164	241	225
dont-know	5	10	11	1	22	0
error	35	15	51	11	5	33

Table 1: Summary of the test datasets and the human labels broken down by judgment category.

> *Imagine you issue a search query to a commercial search engine. Amongst the results set, you see the result below, which someone in your social network has either liked, disliked or shared. Your task is to judge the utility of this social annotation. Value can be assessed on that the annotation is relevant to you, is useful to you, or is generally interesting to you.*

The annotators were also presented with their virtual social network from Figure 2 along with a textual description of each of the twelve individuals in the network. For each test case, the social annotation was graphically presented as though returned from a search engine, similar to what is illustrated in Figure 1. A textual description of the annotation is also presented to the annotator, stating the connection's circle, affinity, expertise, interest valence and geo-distance (if the query was a local query, such as for a restaurant).

For each $\{q, u, c, v\}$ tuple in `HEAD` and `TAIL` the annotation task is to assess the utility of the social annotation according to the following guidelines:

- **Significant utility (`sig-util`)**. The annotation is substantially relevant, useful, or of interest to you.
- **Some utility (`some-util`)**. The annotation is somewhat relevant, useful or of interest to you.
- **No utility (`no-util`)**. The annotation is not relevant, useful or of interest to you.
- **Don't know (`dont-know`)**. I don't have enough information to assess this annotation.
- **Non English/Service error (`error`)**. Can't judge because content is non-English or there is a service error (e.g., 404 message, image didn't load, etc.)

Annotators were encouraged to enter a comment justifying their judgments and were required to do so if their judgment was `dont-know` or `error`.

Each test tuple in `HEAD` and `TAIL` was annotated by two judges randomly drawn from a pool of seven paid professional independent annotators, for a total of 7532 judgments. We trained the judges by iterating on the guidelines over an independent dataset [3]. In cases where the judges disagreed, we adjudicated as follows. If one judge added a comment that we determined clearly justified the judgment, we adjudicated the test tuple to that judge's decision. If both judges added a comment deemed clearly justifying their decision or if both judges omitted a comment, then we retained the disagreeing judgments.

[3]The training phase was necessary to achieve a fair inter-annotator agreement. As such, although it would have been desirable to crowdsource the judgments we deemed the task too difficult and that paid professional independent judges were necessary.

$\mathbf{R}(\mathbf{T})$	ω function definition	
$\mathbf{R_{Rel}}(\mathbf{T})$	$\omega_{Rel} =$ `sig-util:`	1
	`some-util:`	0.5
	`no-util:`	0
$\mathbf{R_{Prec}}(\mathbf{T})$	$\omega_{Prec} =$ `sig-util:`	1
	`some-util:`	1
	`no-util:`	0

Table 2: Utility metrics of social annotations.

Table 1 summarizes the breakdown of the resulting judgments. Inter-annotator agreement as measured by Fleiss' κ on this annotation task was 0.395 (0.487 on `HEAD` and 0.236 on `TAIL`), considered moderate to fair agreement. For our final datasets `HEAD` and `TAIL` we omit the 2.7% of the judgments that were judged as `dont-know` and `error`.

4.4 Utility Analysis

Let T be a set of test tuples, such as those from `HEAD` and `TAIL`, where $t_i = \{q_i, u_i, c_i, v_i\}$, let A be a set of aspect values and T_A be the subset of tuples in T matching an aspect value in A. For example, if $A = \{$`QA-CLS-hea`, `SA-CIR-fam`, `SA-INT-lik`$\}$, then T_A is the set of all test tuples in `HEAD` and `TAIL` where q is a health query, and c is a family connection that has liked u. Finally, let $J(t)$ be the set of judges that annotated a test tuple t.

We define $R(T)$, the expected utility of social annotations in T, as the average utility of each tuple in T:

$$R(T) = \frac{\sum_{t \in T} \frac{\sum_{j \in J(t)} \omega(t,j)}{|J(t)|}}{|T|} \quad (1)$$

where $\omega \colon \mathbb{J} \rightarrow \mathbb{R}$ is a real-valued utility function mapping judgments $\mathbb{J}$ to real numbers in the range [0,1], where $\mathbb{J}$ is the set {`sig-util`, `some-util`, `no-util`} defined in Section 4.3.

Table 2 lists the two variants of $R(T)$ that we report on in this paper. $\mathbf{R_{Rel}}(\mathbf{T})$, our relevance utility metric, assigns a graded utility score to each judgment similar to that done for query-url relevance judgments [16]. $\mathbf{R_{Pre}}(\mathbf{T})$ expresses a binary utility where a social annotation has positive utility if it is judged as either significantly or somewhat relevant, otherwise negative utility. $R_{Pre}(T)$ can be thought of as a measure of social annotation precision.

Table 3 lists the overall utility and per aspect utility breakdown, over the `HEAD` and `TAIL` data sets[4]. The expected overall relevance, at 0.543 indicates that social annotations are generally somewhat relevant, useful or of interest to users. Each individual aspect, however, influences the utility in very different ways. Those aspects with significantly more utility are bolded with a ‡ symbol and those with less utility are bolded with a † symbol.

Overall, we observe no statistically significant difference between the expected utility of head vs. tail queries (note that health queries and restaurant queries seem to have higher utility on tail, but this is not statistically significant). It is surprising that the query class aspects (`QA-CLS`) generally do not show different utility with respect to the average, however further analysis in Section 4.5 reveals significantly differentiated influence in combination with social aspects. Also counter to our expectations, judges found equal utility between navigational and non-navigational queries. For the content aspects (`CA`), although we observe no statistical

[4]`CA-det` had too few judgments to report results.

		HEAD			TAIL		
		Samples	R_{Rel}	R_{Prec}	Samples	R_{Rel}	R_{Prec}
	ALL	2388	0.543 ± 0.011	0.771 ± 0.013	1375	0.541 ± 0.014	0.763 ± 0.016
QA	QA-INT-nav	785	0.539 ± 0.019	0.775 ± 0.022	93	0.513 ± 0.054	0.753 ± 0.066
	QA-INT-nnv	1603	0.545 ± 0.014	0.769 ± 0.015	1282	0.543 ± 0.015	0.764 ± 0.017
	QA-CLS-com	701	0.544 ± 0.021	0.785 ± 0.023	208	0.556 ± 0.035	0.784 ± 0.039
	QA-CLS-hea	169	0.549 ± 0.040	0.772 ± 0.047	36	0.611 ± 0.073	0.847 ± 0.085
	QA-CLS-mov	233	0.526 ± 0.037	0.740 ± 0.041	61	0.549 ± 0.070	0.762 ± 0.077
	QA-CLS-mus	674	0.570 ± 0.020	$\mathbf{0.809^{\ddagger} \pm 0.023}$	762	0.532 ± 0.019	0.755 ± 0.022
	QA-CLS-res	159	0.524 ± 0.045	0.730 ± 0.052	51	0.608 ± 0.073	0.843 ± 0.069
	QA-CLS-oth	452	0.517 ± 0.027	$\mathbf{0.723^{\dagger} \pm 0.030}$	257	0.531 ± 0.034	0.745 ± 0.041
SA	SA-CIR-wkc	818	$\mathbf{0.494^{\dagger} \pm 0.016}$	0.782 ± 0.021	470	$\mathbf{0.498^{\dagger} \pm 0.022}$	0.779 ± 0.027
	SA-CIR-fam	788	$\mathbf{0.654^{\ddagger} \pm 0.019}$	$\mathbf{0.841^{\ddagger} \pm 0.019}$	431	$\mathbf{0.625^{\ddagger} \pm 0.024}$	$\mathbf{0.810^{\ddagger} \pm 0.027}$
	SA-CIR-frn	782	$\mathbf{0.484^{\dagger} \pm 0.021}$	$\mathbf{0.690^{\dagger} \pm 0.025}$	474	0.508 ± 0.026	$\mathbf{0.706^{\dagger} \pm 0.031}$
	SA-AFF-cls	1169	$\mathbf{0.657^{\ddagger} \pm 0.015}$	$\mathbf{0.845^{\ddagger} \pm 0.015}$	689	$\mathbf{0.622^{\ddagger} \pm 0.020}$	$\mathbf{0.804^{\ddagger} \pm 0.021}$
	SA-AFF-dst	1219	$\mathbf{0.434^{\dagger} \pm 0.014}$	$\mathbf{0.701^{\dagger} \pm 0.019}$	686	$\mathbf{0.460^{\dagger} \pm 0.019}$	0.722 ± 0.025
	SA-EXP-exp	1194	$\mathbf{0.635^{\ddagger} \pm 0.016}$	$\mathbf{0.819^{\ddagger} \pm 0.016}$	686	$\mathbf{0.609^{\ddagger} \pm 0.020}$	0.799 ± 0.022
	SA-EXP-nex	1194	$\mathbf{0.451^{\dagger} \pm 0.014}$	$\mathbf{0.723^{\dagger} \pm 0.019}$	689	$\mathbf{0.474^{\dagger} \pm 0.019}$	0.728 ± 0.025
	SA-GEO-nea	57	0.469 ± 0.067	0.711 ± 0.087	13	$\mathbf{0.673^{\ddagger} \pm 0.098}$	$\mathbf{0.923^{\ddagger} \pm 0.098}$
	SA-GEO-far	75	0.593 ± 0.063	0.787 ± 0.070	17	0.588 ± 0.081	0.853 ± 0.108
	SA-GEO-n/a	2256	0.543 ± 0.012	0.772 ± 0.013	1345	0.539 ± 0.015	0.761 ± 0.017
	SA-INT-dis	740	0.570 ± 0.019	$\mathbf{0.818^{\ddagger} \pm 0.021}$	454	0.526 ± 0.024	0.771 ± 0.029
	SA-INT-lik	853	$\mathbf{0.620^{\ddagger} \pm 0.018}$	$\mathbf{0.861^{\ddagger} \pm 0.017}$	498	$\mathbf{0.628^{\ddagger} \pm 0.024}$	$\mathbf{0.827^{\ddagger} \pm 0.024}$
	SA-INT-shr	795	$\mathbf{0.436^{\dagger} \pm 0.019}$	$\mathbf{0.631^{\dagger} \pm 0.024}$	423	$\mathbf{0.455^{\dagger} \pm 0.025}$	$\mathbf{0.680^{\dagger} \pm 0.032}$
CA	CA-per	63	0.603 ± 0.075	0.786 ± 0.075	86	$\mathbf{0.628^{\ddagger} \pm 0.059}$	0.826 ± 0.055
	CA-exc	390	0.561 ± 0.025	0.809 ± 0.028	215	$\mathbf{0.627^{\ddagger} \pm 0.032}$	$\mathbf{0.877^{\ddagger} \pm 0.031}$
	CA-goo	644	0.572 ± 0.021	$\mathbf{0.816^{\ddagger} \pm 0.023}$	337	0.582 ± 0.029	0.812 ± 0.033
	CA-fai	577	0.535 ± 0.023	0.758 ± 0.025	235	0.547 ± 0.032	0.783 ± 0.037
	CA-bad	710	$\mathbf{0.509^{\dagger} \pm 0.021}$	$\mathbf{0.719^{\dagger} \pm 0.025}$	496	$\mathbf{0.458^{\dagger} \pm 0.024}$	$\mathbf{0.660^{\dagger} \pm 0.030}$
	CA-det	4	–	–	6	–	–

Table 3: Utility of social annotation on SERP vs. social relevance aspects from the query, network connection, and content, with 95% confidence intervals. Bold indicates statistical significance over all test tuples (ALL) with † indicating lower utility and ‡ indicating higher utility. (-) indicates too few judgments.

significance within the class, the descending utility trend follows the graded relevance judgments, with higher utility on both HEAD and TAIL for perfect and excellent content versus lower utility for fair and bad content.

The social aspects SA generally have significant differentiating influence. The social affinity (SA-AFF) shows the most influence on utility followed by expertise (SA-EXP) and connection circle (SA-CIR), where colleagues and friends have equal utility but family members have much higher expected utility. For interest valence (SA-INT), knowing that a connection has liked (lik) a link shows more utility than average, but a share (shr) shows a negative utility influence. Interestingly, disliking a link (dis) has little effect on utility, which is further confirmed in our analysis in Section 4.5. Since only 7% of the queries in HEAD and 4% in TAIL were local queries, our sampling resulted in too few geo-distance (SA-GEO) instances that were near or far. The result is large confidence bounds and the only statistically significant utility difference is shown in the TAIL where knowing near-ness is more valuable than not. Further investigation on this aspect is warranted because we expect geographical distance to have influence on social annotations of local queries.

We performed feedback analysis by inspecting a random sample of the comment boxes filled by our judges. For the no-util judgment, the feedback was mainly split between poor content matches or vague queries, and connections of very poor perceived utility. For the latter, expertise was the main discussed social cue followed by affinity and interest valence. Example feedback includes: "Because he is a distant friend, neutral and a non-expert, his opinion is not going to be useful to me." and "Even though the connection is a family member, being a non-expert and neutral on the page would make me think that they do not know a lot on the result." For the some-util judgment, feedback mainly revolved around the circle and dislike aspects as the driving cues for utility. Example comments include: "Chris' dislike might get me to click another link, even though it's what I'm looking for, it could be a bad quality link."; "Even though Bob is neutral, since he's an expert there is added value in his annotation."; and "Anytime one of my friends 'dislikes' a website, it might make me dig further to see why." Finally, for the sig-util judgment, expertise and affinity drove the utility cues. Example feedback includes: "There is a lot of added value to this annotation because Bob is my close friend and an expert." and "She is only a colleague, but she is an expert on the result and her opinion matters to me because she knows what she is talking about."

Finally, we further analyzed the utility of all pairwise combinations of aspects. We list the top-20 and bottom-20 in terms of R_{Rel} utility in Table 4, computed over $T_{\text{HEAD}\cap\text{TAIL}}$. The main conclusion from this analysis is that social annotations have very large variations in utility depending on the aspects defined in Section 4, with relevance utility ranging from 0.336-0.754 in pairwise combinations. These should be leveraged in deciding when to impress a social annotation to a user.

4.5 Aspect Interplay Analysis

We turn now to measuring the interplay between social relevance aspects. In a production system, some aspects are more expensive than others to obtain. For example, most search engines will already have query classifiers in place whereas it may be harder to obtain some of the social aspect values such as a connection's affinity or expertise. In this section, we are interested in answering the question: what is the value of a set of aspects if we know another set of aspects?

Top-20 Pairs by R_{Rel}			
SA-AFF-cls ∩ CA-per	0.754	SA-AFF-cls ∩ CA-exc	0.683
SA-CIR-fam ∩ CA-per	0.736	SA-CIR-fam ∩ CA-goo	0.680
SA-AFF-cls ∩ SA-INT-lik	0.734	SA-AFF-cls ∩ CA-goo	0.679
SA-AFF-cls ∩ SA-EXP-exp	0.731	SA-AFF-cls ∩ SA-GEO-far	0.679
SA-CIR-fam ∩ SA-INT-lik	0.727	SA-EXP-exp ∩ CA-per	0.674
SA-CIR-fam ∩ SA-EXP-exp	0.726	SA-AFF-cls ∩ QA-CLS-res	0.671
SA-EXP-exp ∩ SA-INT-lik	0.713	SA-INT-lik ∩ CA-exc	0.669
SA-INT-lik ∩ CA-per	0.696	SA-GEO-far ∩ SA-INT-lik	0.669
SA-CIR-fam ∩ SA-GEO-far	0.689	SA-CIR-fam ∩ QA-CLS-res	0.668
SA-CIR-fam ∩ CA-exc	0.686	SA-EXP-exp ∩ CA-goo	0.667
Bottom-20 Pairs by R_{Rel}			
SA-CIR-frn ∩ CA-bad	0.422	SA-EXP-nex ∩ SA-GEO-nea	0.396
SA-EXP-nex ∩ CA-bad	0.417	SA-AFF-dst ∩ CA-bad	0.396
SA-INT-shr ∩ QA-CLS-mov	0.412	SA-GEO-nea ∩ CA-fai	0.395
SA-CIR-frn ∩ SA-EXP-nex	0.410	SA-CIR-frn ∩ SA-INT-shr	0.394
SA-CIR-wkc ∩ SA-EXP-nex	0.410	SA-CIR-frn ∩ SA-GEO-nea	0.388
SA-AFF-dst ∩ QA-CLS-mov	0.409	SA-GEO-nea ∩ SA-INT-shr	0.385
SA-CIR-wkc ∩ SA-INT-shr	0.405	SA-AFF-dst ∩ SA-EXP-nex	0.363
SA-INT-shr ∩ CA-bad	0.397	SA-EXP-nex ∩ SA-INT-shr	0.355
SA-EXP-nex ∩ SA-GEO-nea	0.396	SA-AFF-dst ∩ SA-INT-shr	0.355
SA-AFF-dst ∩ CA-bad	0.396	SA-CIR-frn ∩ SA-AFF-dst	0.336

Table 4: Top-20 and Bottom-20 social relevance aspect combinations in terms of R_{Rel} utility.

4.5.1 Relative Gain

More formally, given a corpus of tuples T and a set of aspect values A_1, we are interested in the expected relative gain or loss in social annotation utility, $RG_T(A_2 \parallel A_1)$, if we learn another set of aspect values A_2:

$$\mathbf{RG_T}(A_2 \parallel A_1) = \frac{R(T_{A_1} \cap T_{A_2}) - R(T_{A_1})}{R(T_{A_1})} \quad (2)$$

Consider for example $A_1 = \{\texttt{health}\}$ (query class) and $A_2 = \{\texttt{family}\}$ (circle). Using Eq. 1, we can compute the expected utility of a social annotation in $T_{\{\texttt{health}\}}$ as $R(T_{\{\texttt{health}\}})$. We can similarly compute the expected utility if we also knew the `family` aspect, $R(T_{\{\texttt{health}\}} \cap T_{\{\texttt{family}\}})$. Eq. 2 measures the ratio between these two utilities, which captures the expected gain or loss of knowing `family` given that we knew `health`. RG_T is non symmetric, that is $\mathbf{RG}(T_{A_1} \parallel T_{A_2}) \neq \mathbf{RG}(T_{A_2} \parallel T_{A_1})$. Positive gain is indicated by the sign of RG_T, and no gain is observed if $RG_T = 0$.

Relationship to information gain.

The interplay between social relevance aspects can also be expressed in terms of information gain. Although we report only values of RG_T in this paper and prefer its interpretability (it can be directly interpreted as the expected ratio increase in utility), for completeness we derive the information gain criteria. First, let $P_T(v)$ be the probability that a tuple $t \in T$ is judged as $v \in \mathbb{J}$:

$$\mathbf{P_T}(v) = \frac{\sum_{t \in T} \sum_{j \in J(t)} \phi_j(t, v)}{\sum_{t \in T} \sum_{j \in J(t)} 1} \quad (3)$$

where $\phi_j(t, v)$ indicates if tuple t is judged as v by annotator j. Then the information gain of A_2 given A_1 is defined as:

$$\mathbf{IG_T}(A_1, A_2) = H(T_{A_1}) - H(T_{A_1} \cap T_{A_2}) \quad (4)$$

$$\mathbf{H}(T) = -\sum_{v \in \mathbb{J}} P_T(v) \log_2 P_T(v) \quad (5)$$

4.5.2 Results

We focus our analysis on the relative gain of social aspects with respect to each other, the query class aspects (`QA-CLS`) and content aspects (`CA`). We omit the geo-distance aspect for lack of space and since very few gains were significant.

Table 5 and Table 6 list the relative utilty gains of `SA` vs. `QA-CLS` and `CA`, respectively. Bolded entries indicate statistically significant gains. The value of any cell can be interpreted as the utility gain or loss (in terms of R_{Rel} and R_{Prec}) over both `HEAD` and `TAIL` of knowing the row aspect given the column aspect, i.e. $RG_{\texttt{HEAD} \cup \texttt{TAIL}}(\texttt{row} \parallel \texttt{column})$. For example, given the query class `movie`, knowing that the connection is in the `family` circle increases the relevance utility by a factor of 0.27, whereas if it is in the `work colleague` circle reduces the relevance utility by a factor of 0.132. Dashed entries represent either <0.001 utility gain/loss or impossible combinations (e.g., since family members are always considered to have close affinity in our setting then $T_{\texttt{family} \cap \texttt{distant}}$ is an empty set).

Generally, knowing all social aspects with the exception of `friend` and `dislike` yields significant utility gain (or loss) over the query classes and content relevance aspects. The top-3 social aspects resulting in the most overall gains are the `family` circle and the `affinity` aspects. We computed RG between all combinations of `affinity` and `circle` aspects (table omitted for lack of space). We found that if we knew the `circle` then further knowing `affinity` leads to a utility difference of a factor of 0.319. In contrast, if we knew first the `affinity`, then further knowing the `circle` leads to a utility difference of a factor of only 0.243. Clearly `circle` and `affinity` are not independent. Hence, if one has to choose, investing in obtaining the `affinity` of a connection is more valuable than obtaining `circle`. If the affinity is known to be `close` then there is no value in also knowing that the circle is a `friend` (i.e., a social annotation from a close friend or family member has equal utility in our data). If the affinity is `distant`, however, then there is significant value in determining if the circle is `work colleague` (0.116 gain) or `friend` (0.243 loss).

Expertise, as expected, affects utility relatively more than other social aspects for `health` queries. With respect to `music` queries, `expertise` along with `affinity`, `interest` and the `family` circle equally most influence utility. Content that was `shared` has more influence on utility than other social aspects for `music` queries.

5. PREDICTING SOCIAL RELEVANCE

In the previous sections we have identified and examined the influence of social relevance aspects on the utility of social annotations, and our user study confirmed that this influence is of a differentiated and complex nature. In this section we aim to build on these results by asking the question: can we predict automatically whether a social annotation adds utility to a search result? To address this question, we develop discriminative models by learning from signals obtained in two different ways: (1) **offline** features, obtained from the social relevance aspects used in our user study in Section 4; and (2) **online** features, obtained from signals available at runtime in a commercial search engine, to examine how well we can perform in our prediction task in a real-world scenario without access to features such as a connection's `circle` and `affinity` or gold judgments on `query class` and `content relevance`.

5.1 Offline Features

Our 16 **offline** features are derived from the social relevance aspects presented in Section 3. The query class aspects (`QA-CLS`) are mapped to five binary features, namely `commerce`, `health`, `movie`, `music`, and `restaurant`. The social aspects each become a categorical feature as follows: `circle` = {*wkc*, *fam*, *frn*}; `affinity` = {*cls*, *dst*}; `expertise` = {*exp*, *nex*}; `geo-distance` = {*nea*, *far*, *n/a*}; and

	QA-CLS-com		QA-CLS-hea		QA-CLS-mov		QA-CLS-mus		QA-CLS-res	
	RG_{Rel}	RG_{Prec}	RG_{Rel}	RG_{Prec}	RG_{Rel}	RG_{Prec}	RG_{Rel}	RG_{Prec}	RG_{Rel}	RG_{Prec}
SA-CIR-wkc	**−0.124**	−0.011	−0.029	0.043	−0.132	−0.019	−0.059	0.037	−0.092	−0.011
SA-CIR-fam	**0.215**	**0.093**	0.14	0.056	**0.27**	**0.148**	**0.144**	**0.059**	**0.247**	**0.15**
SA-CIR-frn	−0.078	**−0.074**	−0.081	−0.091	−0.103	−0.12	**−0.086**	**−0.096**	−0.139	−0.132
SA-AFF-cls	**0.211**	**0.089**	0.16	0.063	**0.237**	**0.118**	**0.145**	**0.057**	**0.251**	**0.134**
SA-AFF-dst	**−0.213**	**−0.09**	−0.125	−0.049	**−0.225**	−0.112	**−0.154**	**−0.061**	**−0.208**	−0.111
SA-EXP-exp	**0.159**	0.052	**0.169**	0.062	0.141	0.056	**0.141**	**0.05**	**0.191**	0.075
SA-EXP-nex	**−0.161**	−0.053	**−0.167**	−0.062	−0.126	−0.05	**−0.137**	−0.048	−0.161	−0.063
SA-INT-dis	0.028	0.057	−0.022	0.02	0.044	0.063	0.012	0.029	0.034	0.062
SA-INT-lik	**0.149**	**0.112**	0.146	**0.129**	**0.189**	**0.143**	**0.14**	**0.084**	0.133	0.119
SA-INT-shr	**−0.185**	**−0.17**	−0.134	−0.155	**−0.193**	**−0.174**	**−0.188**	**−0.138**	−0.184	**−0.199**

Table 5: Relative utility gain (R_{Rel} and R_{Prec}) of social aspects (rows) vs. query aspects (columns) over HEAD ∪ TAIL. Bold indicates that the relative utility gain is statistically significant with 95% confidence.

	CA-per		CA-exc		CA-goo		CA-fai		CA-bad	
	RG_{Rel}	RG_{Prec}	RG_{Rel}	RG_{Prec}	RG_{Rel}	RG_{Prec}	RG_{Rel}	RG_{Prec}	RG_{Rel}	RG_{Prec}
SA-CIR-wkc	−0.052	0.042	−0.039	0.041	**−0.093**	−0.014	**−0.102**	0.014	**−0.09**	0.039
SA-CIR-fam	**0.248**	**0.153**	**0.126**	0.056	**0.172**	**0.08**	**0.186**	**0.081**	**0.235**	**0.09**
SA-CIR-frn	−0.107	−0.123	−0.072	**−0.09**	−0.082	−0.064	−0.075	**−0.091**	**−0.151**	**−0.15**
SA-AFF-cls	**0.247**	**0.143**	**0.153**	**0.068**	**0.175**	**0.09**	**0.183**	**0.066**	**0.209**	**0.074**
SA-AFF-dst	**−0.193**	−0.112	**−0.14**	−0.063	**−0.192**	**−0.098**	**−0.189**	−0.068	**−0.188**	−0.067
SA-EXP-exp	0.129	0.044	**0.154**	0.036	**0.159**	**0.063**	**0.151**	**0.061**	**0.145**	0.057
SA-EXP-nex	−0.146	−0.049	**−0.149**	−0.034	**−0.155**	**−0.062**	**−0.174**	**−0.071**	**−0.135**	−0.053
SA-INT-dis	−0.066	−0.004	–	0.046	0.018	0.047	0.026	0.039	0.044	0.043
SA-INT-lik	0.166	0.074	**0.153**	**0.08**	**0.169**	**0.108**	**0.159**	**0.115**	**0.11**	**0.108**
SA-INT-shr	−0.101	−0.072	**−0.16**	**−0.122**	**−0.215**	**−0.176**	**−0.215**	**−0.176**	**−0.158**	**−0.155**

Table 6: Relative utility gain (R_{Rel} and R_{Prec}) of social aspects (rows) vs. content aspects (columns) over HEAD ∪ TAIL. Bold indicates that the relative utility gain is statistically significant with 95% confidence.

interest-valence = {*lik*, *shr*, *dis*}. Finally, the content relevance aspects (CA) are mapped to six binary features, namely perfect, excellent, good, fair, bad, detrimental.

5.2 Online Features

We turn now to features that are available to a search engine at runtime, which we call **online** features. Although no human annotator can provide relevance or query classification judgments at runtime, most search engines today have proxies from automatic query classifiers [21] and content rankers [19]. Social aspects are also generally unavailable at runtime. However, there are a multitude of other measurements computed by search engines, and usage logs are routinely collected. The total number of features we collect in our **online** models runs at over 150. Below we describe the major classes of features and list examples.

- **Query Classes**: Our search engine evaluates each query by a set of automatic query classifiers, which assign a score to each query/class pair. We selected 55 of these classes as real-valued features, including is-commerce, is-local, is-health, and is-movie. While these automatic classifiers are not perfectly accurate, they serve as a proxy for the coarser (but more reliable) query-class annotation in our user study.
- **Session Metrics**: For user sessions containing the query, we measure features such as average session duration and average page view count.
- **User Metrics**: For users that have issued the query, we measure features such as average click count and average page view count.
- **Query Metrics**: We measure query-level aggregate features such as average dwell time, average time to first click, and issues per day.
- **Result Metrics**: We measure query-url aggregate metrics such as average dwell time and abandonment.

5.3 Model

For all our prediction experiments we use the Multiple Additive Regression Trees (MART) [27] algorithm, which is based on the Stochastic Gradient Boosting paradigm [11]. We used log-likelihood as the loss function, steepest-descent as the optimization technique and binary decision trees as the fitting function. MART offers a range of crucial advantages: it has been proven to yield high accuracy, it does not require any feature normalization and can handle any mix of real-valued, multi-valued and binary features, and finally, through its use of decision trees it can handle non-linear dependencies between the features. The latter advantage is of particular importance in our case since we have already found in Section 4 that combinations of social relevance aspects are predictive for social relevance. We cast our task as a supervised learning problem for predicting the utility of a social annotation $t = \{q, u, c, v\}$ (see Section 3). For reasons of simplicity, we cast the prediction task as a binary task (i.e., a social annotation is either relevant or not).

5.4 Experimental Setup

We use the data produced by our user study in Section 4. For each judged social annotation tuple t in the data, we extract a feature vector according to Sections 5.1 and 5.2. The online aggregate features are extracted from US English Web search usage logs from the same 3-month period as the samples drawn for our user study (see Section 4.1). The online features are directly obtained from the annotated tuples from our user study. We retain a total of 2380 HEAD tuples and 1371 TAIL tuples[5].

For each tuple, we mapped the two annotator judgments from the space $\mathbb{J}$ to a binary value indicating whether the social annotation in the tuple was relevant (1) or not (0). To this end, we use a conservative minimum-based approach, where we label any tuple as relevant if the minimum anno-

[5]12 of the 3763 tuples were discarded where online feature extraction failed.

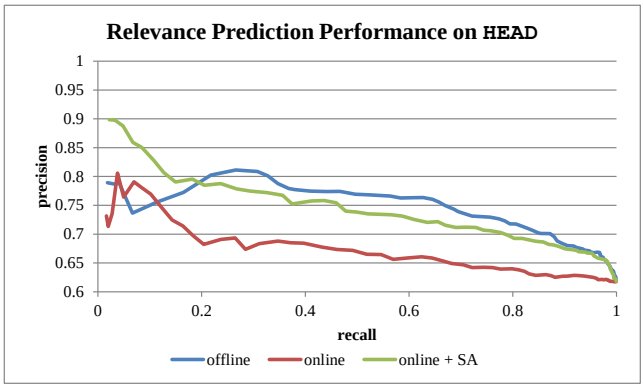

Figure 3: Prediction performance on HEAD using feature sets: offline, online, online + social aspects.

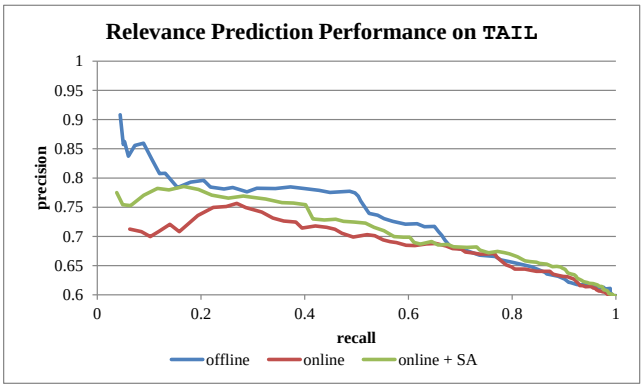

Figure 4: Prediction performance on TAIL using feature sets: offline, online, online + social aspects.

tated utility is `some-util`, i.e. where none of the judges has labeled the case as `no-util`. In HEAD this results in 909 class 0 (non-relevant) examples and 1471 class 1 (relevant) examples. In TAIL we have 552 non-relevant and 819 relevant examples.

For the MART learner, we train for 100 iterations (resulting in 100 trees) and restrict the decision tree stumps to 10 leaf nodes that cover a minimum of 25 samples. All reported results are based on 10-fold cross-validation.

5.5 Results

We present the results of our prediction experiments separately for HEAD and TAIL tuples since we observe systematically different behavior between the two in our prediction task. Figures 3 and 4 illustrate the precision-recall characteristics of the relevance prediction with different feature sets. Overall, the task is learnable and as could be expected, the prediction model that uses **offline** features outperforms all other models on both HEAD and TAIL.

It is common in practice to impress a social annotation anytime one is available for a query-url context. Looking at the split between positive and negative examples in the test data, one would achieve 61.8% relevance precision on HEAD and 59.7% relevance precision on TAIL by impressing every available social annotation to a user. Using our **offline** model on HEAD, one could increase the rate of relevance by a factor of 13% while maintaining 88% recall, or by a factor of 25% at 50% recall. On TAIL, the same model would increase the rate of relevance by a factor of 7% at 87% recall, or by a factor of 29% at 50% recall.

We analyzed the importance of the offline features in our model by observing the weights assigned by MART in its training log files. For HEAD, the social aspects were consistently ranked highest: circle, affinity and expertise are the three most important features, followed by content relevance judgments. For TAIL, the picture is different. Here, the content relevance features are dominant. The top-ranked feature is whether the content relevance is marked as `bad`, which is not surprising since instances of non-relevant urls are higher in TAIL and social annotations on irrelevant documents have less utility. This relevance feature is followed in the importance ranking by a number of social aspects (affinity, circle and expertise), followed by whether the content relevance is `excellent`, followed by query class features.

We also observe that there is enough signal in the **online** features for predictions up into the 70-75% precision range, albeit at much lower recall rate than for offline features.

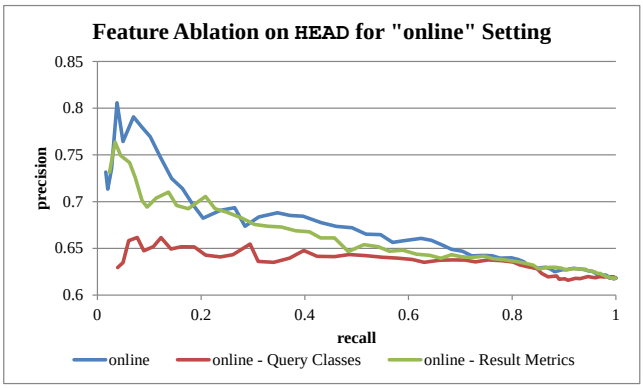

Figure 5: Ablation of two most predictive online feature familes on HEAD: Query Classes and Results Metrics.

Prediction ability is highest on TAIL nearly achieving the performance of offline features. Somewhat surprisingly, in both HEAD and TAIL, we observed that the **Query Class** features were by far the most predictive, followed by the **Results Metrics** features - see ablation results in Figure 5 for HEAD (TAIL omitted for lack of space). In fact for HEAD queries, only using query class features produces results close to those using all online features. In TAIL, the picture is more differentiated, where adding other online features to the query classifier features improves results.

Based on this analysis of feature importance and the finding that online features are predictive but not as predictive as offline features, we also experimented with adding the most predictive offline features (aspect family) to the set of online features. For HEAD queries we added the social aspect features (`circle`, `affinity`, `expertise`, etc.), and for TAIL queries we added the content relevance features (`perfect`, `excellent`, `fair`, etc.) Results are also shown in Figures 3 and 4. We observe that we can increase the performance of online features by adding social aspects and content relevance features. We take this as an encouraging result since proxies of both these feature families may be made available at runtime. Content relevance features such as PageRank scores can be assessed, which - while not as accurate as human judgments - may provide additional signal. Similarly, social relevance features could be amenable to statistical modeling from observable data such as social network characteristics or profile modeling at runtime, an area that we plan to investigate in future research.

6. CONCLUSION AND FUTURE WORK

We presented a taxonomy of aspects that influence the perceived utility of social annotations in a Web search scenario, drawn from the query, social connection, and content relevance. Via a user study, we took a first step at quantifying the utility of social annotations and gained insights on the complex interplay between the social relevance aspects. We concluded that there are large variations in utility depending on the aspects in play and that these should be leveraged when deciding to impress a social annotation to a user. We learned that social aspects are most influential in perceived utility, in particular affinity, expertise and interest valence. We further established that close social connections and experts in the search topic provide the most utility, whereas distant friends and friends that show no positive or negative interest valence provide the least utility, by a factor of over 50%.

We also showed that we can automatically predict whether a social annotation is relevant for a given query/url pair. We cast the task as a binary supervised learning problem over a stochastic gradient boosting model. In an offline experiment, we drew features from the manually labeled user study and established that we can accurately predict social utility. In a configuration simulating an online scenario, we drew session-, query-, document-, and user-level features from query classifiers and web usage logs, which can be computed at runtime by commercial web search engines. In this online setting, we established the prediction task as learnable and approaching the performance of the offline model. Finally, by adding the social aspects and content relevance aspects from the offline features to the online features, we gained predictive performance over just the online features.

A promising avenue of future work is to develop social aspects classifiers in order to increase our ability to predict the utility of a social annotation. Other directions include investigating the influence of other social aspects such as age, gender, user location, and school; and applying our framework to other social features such as interleaved results. Perhaps most valuable, however, is to broaden our concept of utility, which we have limited to a search result, to the whole-page user experience, as outlined in Section 1, and further our understanding of how social features affect the overall information seeking, discovery, and sensemaking processes.

7. REFERENCES

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *SIGIR*, 2006.

[2] S. Bao, G. Xue, and X. e. a. Wu. Optimizing web search using social annotations. In *WWW*, 2007.

[3] C. L. Barry and L. Schamber. Users' criteria for relevance evaluation: A cross-situational comparison. *Inf. Process. Manage.*, 34(2-3):219–236, 1998.

[4] P. Borlund. The concept of relevance in ir. *JASIST*, 54(10):913–925, 2003.

[5] C. J. C. Burges, K. M. Svore, P. N. Bennett, A. Pastusiak, and Q. Wu. Learning to rank using an ensemble of lambda-gradient models. *Journal of Machine Learning Research*, 14:25–35, 2011.

[6] D. Carmel, N. Zwerdling, I. Guy, S. Ofek-koifman, N. Har'el, I. Ronen, E. Uziel, S. Yogev, and S. Chernov. Personalized social search based on the user's social network. In *CIKM*, pages 1227–1236, 2009.

[7] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *CIKM*, pages 621–630, 2009.

[8] C. Clarke, E. Agichtein, S. Dumais, and R. White. The influence of caption features on clickthrough patterns in web search. In *SIGIR*, 2007.

[9] B. M. Evans and E. H. Chi. An elaborated model of social search. *Inf. Process. Manage.*, 46(6):656–678, 2010.

[10] R. Fidel and M. Crandall. Users' perception of the performance of a filtering system. In *SIGIR*, 1997.

[11] J. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29, 2001.

[12] K. Haas, P. Mika, P. Tarjan, and R. Blanco. Enhanced results for web search. In *SIGIR*, 2011.

[13] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search. In *WSDM*, 2008.

[14] R. A. Hill and R. I. M. Dunbar. Social network size in humans. *Human Nature*, 14(1):53–72, 2003.

[15] S. Huffman and M. Hochster. How well does result relevance predict session satisfaction? In *SIGIR*, 2007.

[16] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20:422–446, October 2002.

[17] T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3:225–331, March 2009.

[18] A. Muralidharan, Z. Gyongyi, and E. H. Chi. Social annotations in web search. In *SIGCHI*, 2012.

[19] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *WWW*, 1998.

[20] J. Pitkow and H. e. a. SchÂÍutze. Personalized search. In *ACM*, volume 45(9), 2002.

[21] D. Shen, J. Sun, Q. Yang, and Z. Chen. Building bridges for web query classification. In *Proceedings of SIGIR*, 2006.

[22] J. Teevan, S. Dumais, and D. Liebling. To personalize or not to personalize: Modeling queries with variation in user intent. In *SIGIR*, 2008.

[23] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *SIGIR*, pages 449–456, 2005.

[24] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. The anatomy of the facebook social graph. *CoRR*, abs/1111.4503, 2011.

[25] K. Wang, T. Walker, and Z. Zheng. Pskip: estimating relevance ranking quality from web search clickthrough data. In *SIGKDD*, 2009.

[26] S. Wedig and O. Madnani. A large-scale analysis of query logs for assessing personalization opportunities. In *SIGKDD*, 2006.

[27] Q. Wu, C. J. Burges, K. M. Svore, and J. Gao. Ranking, boosting and model adaptation. *Microsoft Research Technical Report*, 2008.