Enhanced Results for Web Search

Kevin Haas
Microsoft
1065 La Avenida
Mountain View, CA 94043, US
kevin.haas@microsoft.com

Peter Mika
Yahoo! Research
Diagonal 177
08018 Barcelona, Spain
pmika@yahoo-inc.com

Paul Tarjan
Facebook
1601 S California Ave
Palo Alto, CA 94304, US
pt@fb.com

Roi Blanco
Yahoo! Research
Diagonal 177
08018 Barcelona, Spain
roi@yahoo-inc.com

ABSTRACT

"Ten blue links" have defined web search results for the last fifteen years – snippets of text combined with document titles and URLs. In this paper, we establish the notion of *enhanced search results* that extend web search results to include multimedia objects such as images and video, intent-specific key value pairs, and elements that allow the user to interact with the contents of a web page directly from the search results page. We show that users express a preference for enhanced results both explicitly, and when observed in their search behavior. We also demonstrate the effectiveness of enhanced results in helping users to assess the relevance of search results. Lastly, we show that we can efficiently generate enhanced results to cover a significant fraction of search result pages.

Categories and Subject Descriptors

H.3.3 [**Information Systems Applications**]: Information Search and Retrieval

General Terms

Design, Human Factors

Keywords

search results, user interfaces, web search, semantic web

1. INTRODUCTION

Ten blue links, along with document titles and approximately 100-character summaries (known as abstracts) have dominated Web search results for the last fifteen years. The goal of the abstract is to help users make a quick assessment of whether the Web document returned as a result is relevant to their query or not (see Figure 1). Extensive research has focused on effectively selecting the relevant parts of documents (the snippets) and generating a readable presentation (the abstract) in an efficient manner, e.g. [17, 18, 19, 21, 28, 30]. However, little has changed with the presentation of the search results and the type of data provided within the abstracts.

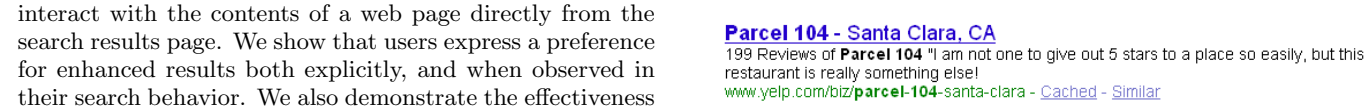

Figure 1: Example of an traditional search result

In this paper, we present a novel way of generating search abstracts based on structured data or *metadata* associated with Web documents. Structured representations of document content allow us to give more relevant and more compelling representations of search results, as well as to imbue them with entirely new interactive functionality. In short, a deeper understanding of Web content enables us to go well beyond the original concept of textual summarization of the document. This deeper understanding is enabled by two major streams of research: Information Extraction (IE) and the Semantic Web.

By now, the Semantic Web has reached a sufficient level of maturity in terms of defining standards (such as RDFa [1] and microformat markup[1]) that publishers can rely on to embed metadata in web pages. In addition, advances in automating extraction, element formalization within the HTML specification, commonalities between Web content management systems, and the increasing importance of head sites within Web search results (from the consumer perspective) allows search engines to leverage Information Extraction technologies at a wider scale. This enables the efficient extraction of structured data from Web documents that otherwise contain no semantic information about the content presented.

In our current work, we exploit both of these methods to go significantly beyond the expressive power of current textual abstracts, and propose the usage of *enhanced search results* that incorporate images, links, key-value pairs and

[1] `www.microformats.org`

SIGIR'11, July 24–28, 2011, Beijing, China.

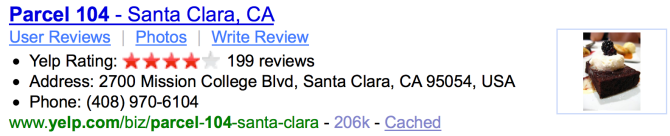

Figure 2: Example of an enhanced search result

interactive elements (see Figure 2). We implemented metadata extraction in our web crawler to support information extraction from Web documents and to collect Semantic Web data necessary to generate enhanced results. Further, we created a system to efficiently develop, manage and execute the computer code necessary to translate structured data into search result presentations. This system is currently in production as part of our Web search engine where at least half of all search result pages currently served (October, 2010) contain at least one enhanced result. We evaluate our work by addressing the two foremost questions about enhanced search results.

- Do users have a preference for enhanced search results and are they helping them to find what they are looking for?
- Does the method for generating enhanced search results scale to the Web in terms of content coverage and user impressions?

Based on a side-by-side editorial evaluation and a comparative analysis of click-through rates on a fraction of real search traffic, we show that users both explicitly state their preference for enhanced results in a test scenario and find them more appealing in day-to-day search engine usage. We also measure the usefulness of enhanced results in helping users to distinguish relevant results from irrelevant ones, a crucial benefit for both publishers and search engine providers. Lastly, we show that generating enhanced results scales to a Web search scenario in terms of both content and query coverage.

2. MOTIVATION

The Web was originally designed for presenting information for human consumption. As a result, even though the website owners may put significant effort into modeling their data and managing it in structured forms such as relational tables, OO models or XML schemas, this structure is lost when formatted as text for human consumption. The Semantic Web provides the required standards for sharing structured data across the Web using a generic data model, the Resource Description Framework, or RDF, and for describing the schema of the data using expressive, logic-based languages such as the Web Ontology Language, or OWL.

The Semantic Web is increasingly being adopted in two settings. The Linking Open Data (LOD)[2] community projects use Semantic Web technology to directly expose public data sets in RDF, often for data that has previously not been accessible from the Web. The second main area of attention is the annotation of existing Web resources, in particular HTML documents. This appeals to the large majority of existing publishers who generate their web pages automatically and from an existing structured data source, since publishing data this way requires only minor modifications to the template that generates their pages. From the perspective of a Web search engine, this second area is also the one that holds immediate promise as it is closer to the traditional expectations of search engine processing. Collecting metadata embedded inside HTML pages requires minimal changes to the existing crawling infrastructure, whereas collecting Linked Data requires not only new crawling procedures but also additional mechanisms to establish and transfer trust from the network of HTML documents to the network of RDF data (or vice versa).[3] Allowing independent parties to provide metadata for sites which they do not own could have negative consequences, e.g. a manufacturer making statements about the price of a competing product.

There are several standards and conventions for embedding metadata inside (X)HTML documents. *RDFa* is a W3C recommendation for embedding RDF models inside (X)HTML documents as additional markup. *Microformats* are informally specified social conventions of encoding metadata about particular types of objects, e.g. persons or events. Loosely defined specifications make it easier to annotate using microformats, but microformats on the Web exhibit a more diverse application of the proposed syntax than RDFa.

We implemented support for RDFa and several popular microformats inside the indexer component of Yahoo Search, a major Web search engine. Figure 3 shows the growth in the percentage of indexed URLs with either RDFa or commonly supported microformats[4] in a collection of 12 billion Web documents.

We have argued in our previous work [23] that while similar figures are interesting to observe, what matters is whether this data is useful to satisfy the information needs of search users. In particular, we are interested in how often specific metadata formats would be surfaced by the search engine. To establish this, we replayed a random sample of 7117 queries of one month of Yahoo US query log and observed the returned results (see Table 1).[5] The last two columns show the total number of potential enhanced results and the average number of potential enhanced results per query within the top 10 results.

Based on this analysis, we know that despite having a relatively low percentage of URLs with metadata, on average every search result page features at least one result with page-based metadata, with the most popular hcard microformat returned once in every other search result page. By contrasting Table 1 and Figure 3 we can also see that our expectation is confirmed in that the most popular microformats (e.g. tagging) are not necessarily the ones appearing on pages that are most searched for.

We may note that not all forms of data are equally useful for presenting to the user, including the most popular tagging microformat originally invented for giving hints to the Technorati search engine for categorizing blog posts. RDFa

[2] `www.linkeddata.org`

[3] As an example, the website `http://dbpedia.org` provides RDF metadata that originates (and is extracted from) the website `http://www.wikipedia.org`. However, no formal connection exists between these domains in terms of ownership.

[4] with the exception of fb-img, which we denote as the Facebook Share format (`www.facebook.com/facebook-widgets/share.php`)

[5] The query log was sampled in January, 2009 and the experiment was run in January, 2010.

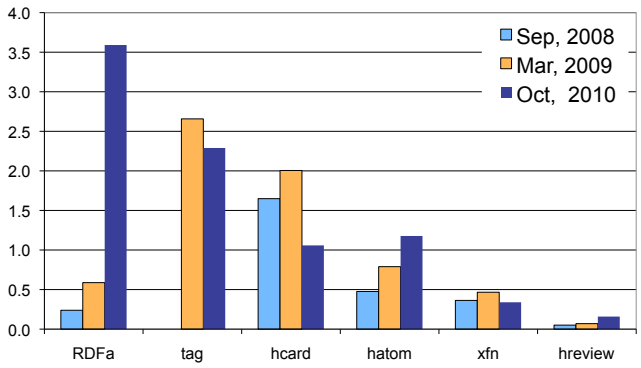

Figure 3: Percentage of Web pages with RDFa data and selected microformats.

Table 1: The number of queries that return zero to five results with various metadata within the first page of search results

format	0	1	2	3	4	5	Sum	Avg
hcard	4729	1702	483	145	38	8	3374	47.4%
license	5773	1294	47	3	0	0	1397	19.6%
adr	6090	714	193	72	19	9	1566	22.0%
hatom	6330	687	89	8	2	1	902	12.7%
fb-img	6379	640	76	14	6	0	872	12.3%
RDFa	6618	462	27	3	2	3	561	7.9%
xfn	6618	462	31	4	1	0	546	7.7%
geo	6716	351	46	4	0	0	455	6.4%
tag	6721	372	24	0	0	0	420	5.9%
ANY	2815	2175	1153	561	249	100	8072	113%

data itself contains information using a number of common and less common ontologies, making it hard to exploit efficiently. We will return to question of how much of this data can be successfully exploited in Section 5.2.

We also note that there are good reasons to move away from an average case analysis. Not only the amount of metadata varies by query category, but we may also attribute different value to different categories of queries, e.g. attributing smaller weight in evaluation to queries that are already "solved", e.g. navigational queries with near perfect results. One might also weight queries based on the possibility for monetization, but also by the value attributed by the users for solving the query. We leave these analysis for future work.

Our analysis above puts an upper bound on exploiting enhanced search results based on embedded metadata alone. The question then is how one could effectively increase the amount of data available for enhanced search results by information extraction. Despite the popular notion of the long tail of Web content, another set of analysis shows us that preferential treatment to "top sites" provides the opportunity for enhanced search results without needing metadata from the long tail. By running the same set of queries through the search engine, we observed that a relatively small number of sites have a disproportionate appearance on result pages compared to their size. For example, Wikipedia appeared in roughly 25% of all search results. Correlating this data with Table 2[6], we can see that despite its relatively small size Wikipedia alone is surfaced more often than even the most common microformats. These results tell us that one could effectively index the data of these top sites by creating wrappers, a fact that we exploit in our work. Such wrappers can be effectively hand-written by website owners or learned through wrapper induction from training data [20].

Table 2: The number of queries that return 0 to 2 results with pages from the top host names

host name	0	1	2	Sum	Avg
en.wikipedia.org	5321	1795	2	1797	25.2%
youtube.com	6691	426	0	426	6.0%
answers.com	6714	403	0	403	5.7%
amazon.com	6722	395	0	395	5.6%
local.yahoo.com	6739	378	0	378	5.3%
blog.360.yahoo.com	6846	271	0	271	3.8%
facebook.com	6867	250	0	250	3.5%
technorati.com	6883	234	0	234	3.3%
ehow.com	6889	228	0	228	3.2%

3. RELATED WORK

3.1 Search result summarization

Traditionally, Web search engines display a list of captions containing a title, an abstract and a URL. The importance of Web search result displays in helping users to determine the relevance of search results is firmly established. Clarke et al [6] explore how caption features influence Web search behavior. Their findings suggest that relatively simple text features like the readability of the snippet, and the length of the URL shown in the caption, can significantly influence users' Web search behavior.

Most improvements to search result presentation have come from advances in text summarization. Varadarajan and Hristidis [31] generate more relevant snippets from spanning trees built from a document graph and compare them to the ones generated by Google and MSN desktop search. Kanungo and Orr [19] describe a machine learned model for predicting the readability of search results by combining text readability measures and other features specific for Web search results (such as the presence of ellipses). With the exception of Cutrell and Guan [7] who investigated the influence of snippet length on Web search peformance using an eye-tracking study, most researchers in this area base their evaluation on user studies of some form. Although various measures exist for assessing the readability of natural language texts, e.g. the classic SMOG measure [22], these are not directly applicable to short runs of incomplete text. For example, in the above mentioned study, Kanungo and Orr train and test their model using human assessments of Web search results, where the judges evaluate readability on a 1-5 scale, from "Unreadable" to "Easy to read". They note that perfect agreement is fairly low (46.4%), though there is a near perfect agreement in 84.5% of the cases. There is no

[6]Note that in general Web search engines hide multiple results from the same host, unless the user explicitly triggers host-based search

standard user evaluation and a number of variations exist in the literature.

Few works have focused on improving the accessibility of Web search results by changing the form of search result displays. Dumais et al. [11] found that users perform search tasks faster if results are grouped by category in the interface. White et al. [33] evaluate techniques to promote user interaction with search results. In their work they treat the amount of time a user spends viewing a summary as an indicator of relevance. They presented three different solutions, based on the display of a summary, list of top-ranked relevant sentences of the results, and dynamic updates of the search result list using implicit feedback.

In our work, we propose new types of search result displays that represent a significant departure from the classical search abstracts in current Web search engines and therefore our first concern is user acceptance. For this reason, we first perform a user study to establish the users' choice for enhanced results using a binary, side-by-side evaluation that is easy to perform and results in higher agreement. We also perform a so-called bucket test, where we expose a fraction of our search traffic to enhanced results over an extended period of time and compare the user's interaction with the baseline. For enhanced results, we show an increase in Click-Through Rate (CTR), a metric commonly used in online advertizing, content and search engine optimization.

Despite the wide appeal of the CTR metric, more sophisticated models are required to consider effects such as the position bias, i.e. that a user may not get to examine certain search results. User interaction models have been developed in the past to explain observed user behavior in query logs [13, 3]. Interaction models have direct applications to ranking as a complement to editorial data but can also be used in comparative search engine evaluation [12, 16]. In our work, we extend Chapelle and Zhang's click model [3] and fit it to large-scale query log data to show that enhanced results are effectively helping users to distinguish relevant results from irrelevant ones. While clicks-through data has been used in improving text-based search result generation [28], to our knowledge this is the first time that interaction models are used in evaluating search result summaries.

3.2 Semantic Search

A number of end-to-end semantic search systems have been developed up to date, e.g. [10, 8, 4, 15, 24, 29, 32], although all of them are operated as academic demos and prototypes. Unlike web search engines, most of them focus on retrieving RDF resources directly instead of textual documents, while some engines perform hybrid retrieval on a collection of documents annotated with metadata, e.g. [34, 14]. With the exception of [24], semantic search engines that perform their own crawling only retrieve data published directly as RDF, and do not extract RDFa or microformats.

Although there is some experimentation in novel displays for Semantic Web search engines, publications on generating snippets in this context are few and far apart. In a Web search setting, Bai et al. worked on snippet generation for a semantic search engine (Sindice) that indexes instance data [2]. They divide the abstract in two parts: the first, static part showing statements related to the main topic of the document, and weighted by the importance of the predicate of the triple, while the second, dynamic part shows statements ranked by their relevance to the query. The evaluation considers the performance of the system, and compares various methods to identify the main topic of documents. Penin et al. [25] operate in the setting of a Semantic Web search engine that indexes only ontologies, and thus consider snippet generation as an ontology summarization problem. They posit that a succinct summary of the ontology should cover all main topics mentioned, and thus perform clustering on the RDF graph, selecting the most relevant statements from each cluster. Their evaluation centers around the effectiveness of clustering and performance, and provide only anecdotal evidence from end users. As in the case of text search results, these authors operating in the semantic search field do not have a standard evaluation methodology or benchmark to rely on.

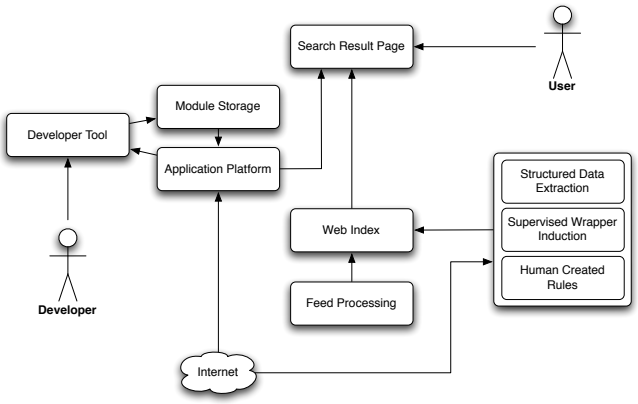

Figure 4: System Architecture

4. IMPLEMENTATION

Enhanced results can be implemented as an extension to a conventional inverted index based search engine, as shown in Figure 4.

In the offline indexing phase, metadata is extracted from Web documents by executing RDFa and microformat extraction, hand-written and machine-learned rules. The difference in RDFa and microformat extraction is that RDFa extraction needs to be implemented once, independent of the schema of the data, while microformat extraction needs to be coded separately for each format. In addition to the information coming from the indexed documents via extraction and markup, our system also admits data feeds submitted by website owners. Due to the potential exposure to spam injection, we allow only the website owner to provide annotations for a given page. As there is no standard format for RDF feeds, we developed our own format for this purpose [7]. Feeds are often preferred by publishers who would not like to make metadata publicly available.

All of the gathered data is stored internally in an RDF-based format. RDF was chosen over other options (such as XML) because of its flexibility: RDF can represent any graph structure, the data and the schema can be stored separately or intermixed, and publishers are free to combine multiple vocabularies if a single vocabulary doesn't match their needs.

[7] http://developer.yahoo.com/searchmonkey/smguide/datarss.html

When a user performs a query, the regular Web search engine retrieval process runs, returning the "ten blue links" and whatever structured data that was associated with them. For the pages that contain metadata, the search engine triggers one of the applicable plugins that can transform the data into the presentation. There are both default plug-ins enabled for all users, and plug-ins developed by external contributors, which require opting in. Note that we also call default plug-ins "object templates", because they only require the presence of particular data to appear, i.e. the data that gives the values for the presentation template. Such object templates exist for a number of popular objects types such business listings, products, video etc.

External plugins take precedence over default plugins. Further, the end user can add or remove both types of plugins via the search engine interface, i.e. there is a possibility for personalization. In order to develop new plug-ins, we have made available an online tool where a developer can

- Define which URLs they would like to change the appearance of (as a wildcard pattern)
- Optionally write custom extraction rules (in XSLT) to pull out any data that wasn't stored in the index already
- Select the sources of data that need to be present for the plug-in to trigger
- Create a translation from data to presentation in PHP
- Choose to list the application in the Application Gallery for other users to use

Although it would have been an option to define the presentation plug-ins using an RDF visual vocabulary such as Fresnel [26], we have opted for a solution where the transformation from data to presentation is described in a controlled subset of PHP. The advantage of choosing a fully-fledged programming language over a declarative mapping is that if/then checks and simple manipulations, such as converting number or date formats, can be still performed at this stage. In PHP, the data is queried by executing queries in a simple path expression language similar to Fresnel path expressions [26], which is more appropriate for this limited task of data selection then the SPARQL standard [27]. In particular, joins are not required. When developing any code in the developer tool, the results of the extraction or translation are shown immediately in a preview to help developers debug their code.

5. EVALUATION

In the evaluation of our work, we seek to answer the two basic questions we asked in Section 1: whether enhanced results bring clear benefits to users, and whether the method of generating them provides sufficient coverage of search results.

5.1 Explicit and Implicit User Feedback

To evaluate the effectiveness of enhanced results, we have first performed a user study to measure the explicit preference of users for these new types of results. Prior to deployment, we have also carried out a large-scale online evaluation using real search engine traffic, designed to measure implicit preference for enhanced results in terms of click-through. Lastly, by modeling user behavior using historical query-log data we show that enhanced results effectively help users to determine the relevance of search results.

5.1.1 Side-by-side editorial evaluation

As put forward in Section 3.1, no gold standard dataset or evaluation methodology exists to assess the quality of search abstracts. However, we can rely on a baseline for comparison, i.e. the existing presentation of search results. Although the methods of generating search result summaries vary by search engine, the appearance of search results are surprisingly uniform.

Using internal editorial resources, we ran side-by-side evaluations of enhanced search result templates and the traditional search result templates containing only textual summaries. Each template evaluation was performed independently using editorial teams ranging from three to fourteen assessors, depending on the number of results being reviewed. Each result pair was evaluated by three editors, and the overall preference was taken as the majority of the individual preferences.

Table 3 contains a subset of the data captured during these editorial reviews. Specific to the product template evaluation, we took 658 commonly-occurring URLs matching our "product" presentation template and asked the editorial team to evaluate the enhanced search result presentation versus the traditional presentation. Each editor was asked to judge which result was better as well as categorize the reason; additional comments could be recorded as feedback to the user experience teams.

Aggregating the data captured in Table 3 allows us to generate Table 4. The majority of results were deemed "better" by the reviewers when presented with an enhanced search result. When the traditional results were preferred, the most common complaints were the incorrectness of the data presented (e.g. incorrect prices) or irrelevance of images.

Table 4: Side-by-side product template analysis

Type	Enhanced	Traditional	No judgment
Preference	84%	3%	13%

5.1.2 Click-through rate analysis

In addition to our editorial user study, we performed a bucket test, a form of online experiment where a small fraction of search sessions is diverted to an alternate version of the search result page during a test period. In our case, the control group was shown a traditional, textual search result for all pages, and the test group was shown an enhanced local search result with the business's telephone number, address, and "curbside image" for websites with local metadata, with all other variables being constant. We used a 5% bucket of Yahoo Search search traffic in the United States and ran the experiment over a period of three months. Note that the sessions, and correspondingly the users are selected randomly from all sessions and users of the search engine, and therefore the query load is also representative of the total query load of the search engine.

We instrumented our search engine to collect CTR (click-through rate) data during the bucket test. CTR is a widely reported metric in search, content optimization and online

Template	URL	A/B	Reason	Comments
Product	#0	A	traditional doesn't convey as much info	
Product	#1	A	traditional doesn't convey as much info	slight preference to A because price is included in the result
Product	#2	-		inclusion of image does not help this result
Product	#3	A	traditional doesn't convey as much info	the price gives A the edge. The image is not particularly compelling.
...	...	...	...	...
Product	#657	B	bad data extraction	extracted price shown on SRP is wrong

Table 3: Subset of side-by-side results for the product template

advertizing. One important consideration in CTR measurement is being able to differentiate a "good click" from a "bad click." Good clicks include no further clicks from that search result or a long dwell time on the target page [9]. We considered a long dwell a time of 100 seconds or longer based on the results of the authors, who show that good clicks measured this way generally indicate a relevant result for the user, and a bad click indicates the search engine surfaced an irrelevant result. Our measure of CTR is thus defined as:

$$CTR = \frac{\#\text{ good clicks on the result}}{\#\text{ total views of the result}}$$

In Table 5, we show a subset of the data collected during this experiment.[8]

Table 5: Subset of bucket CTR data captured

Date	Template	Site	Bucket CTR
t0	Local	site 0	4.65%
t1	Local	site 0	4.97%
t2	Local	site 0	4.70%
t3	Local	site 0	4.54%
t4	Local	site 0	4.52%
...	...	...	...
t100	Person	site n	5.65%

We then generated the aggregation of the data by site and by template for the bucket period, as shown in Table 6.

Table 6: Bucket CTR analysis

Site	Template	Bucket CTR	Control CTR	Change
site 0	Local	5.3%	4.5%	17.8%
site 1	Local	5.8%	5.5%	5.8%
...	...	...	...	...
site n	Person	6.1%	6.0%	16.7%
ALL	ALL	4.6%	4.0%	15.0%

Comparing the CTRs we noticed that the enhanced search results showed an average increase of 15% CTR for all sites, with a maximum of 33% for an enhanced search result for a particular site. We consider this as a positive result, as negative rates might indicate that the result contained exactly the information that the user needed, and the user would not click through to the website itself. While great for a "task complete" metric, website owners would be less than happy with a search engine innovation that used data from their websites but reduced their referrals and the corresponding revenue opportunities.

[8]The data in Table 5 has been partly hidden, but we include this table as a representation of the data captured during the experiment.

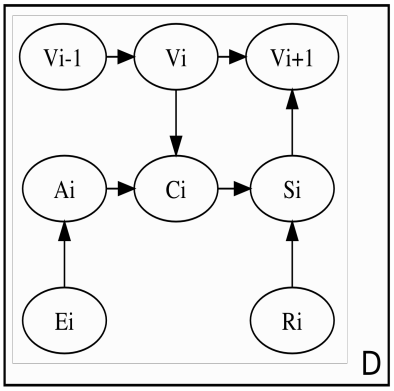

Figure 5: Click model

5.1.3 User interaction model

Click-through rate analysis reveals an implicit preference for enhanced results whenever such a result is displayed, but it can not show whether enhanced results are effective in helping the user to distinguish relevant results from irrelevant ones at the time of deciding on a click. In order to model how enhanced results impact user interaction, we extend Chapelle and Zhang's state-of-the-art click model [3] with the potential influence of enhancing a search result.

The resulting Dynamic Bayesian Network model is presented in Figure 5. The variables inside the plate are defined at a query-session level and define the possible interactions of the user. Snippets are placed in positions from 1 to 10 (denoted by i). All random variables are binary. C_i indicates whether there is a click at position i, V_i if the user has viewed the result, A_i if the user has been attracted by the result, S_i if the user was satisfied with the landing page, E_i indicates whether the result is an enhanced result or not, and R_i the relevance status of the document. The model is further characterized by the following equations [3]:

- $A_i = 1, V_i = 1 \Leftrightarrow C_i = 1$
- $C_i = 0 \Rightarrow S_i = 0$
- $S_i = 1 \Rightarrow V_{i+1} = 0$
- $p(V_{i+1}|V_i = 1, S_i = 0) = \gamma$
- $V_i = 0 \Rightarrow V_{i+1} = 0$

In the original model, only C_i is an observed variable, and R_i is estimated – in our case we will employ true relevance data from those URLs, and therefore R_i is an observed variable. Whether there is an enhanced result or not is also observed in our model, whereas the rest of the variables will be estimated by taking $\gamma = 0.9$, the value which provided the best estimation of relevance in [3]. We also assume that $C_i = 1, R_i = 1 \Rightarrow S_i = 1$.

In words, the assumption of the model is that upon submission of a query, there is a list of results displayed to a user (D). A user scans the result list from top to bottom, examining all the results in order. After viewing a result ($V_i = 1$), a user may get *attracted* by a result ($A_i = 1$), deciding to click on it ($C_i = 1$); if the user clicks on a result but she is not satisfied by the landing document, she examines the next result with probability γ. If the document is relevant we assume the user is satisfied and she stops the search process for the query ($V_{i+1} = 0$)

This model allows us to answer the following questions by probability estimation:

- What is the probability that an enhanced result is relevant *a priori*? That is, $p(R_i = 1|E_i = 1)$ compared to $p(R_i = 1|E_i = 0)$
- What is the probability that an enhanced result viewed by a user is relevant? That is, $p(A_i = 1|E_i = 1) = p(R_i = 1|V_i = 1, E_i = 1)$ compared to $p(R_i = 1|V_i = 1, E_i = 0)$
- What is the attractiveness of enhanced result when viewed by a user? That is, $p(C_i = 1|V_i = 1, E_i = 1)$ compared to $p(C_i = 1|V_i = 1, E_i = 0)$
- What is the probability that the enhanced result was useful? This is, $p(R_i = 1|C_i = 1, E_i = 1)$ compared to $p(R_i = 1|C_i = 1, E_i = 0)$

To compute the above transition probabilities, we took a sub-sample of 10 consecutive days of Yahoo US query log from October, 2010, collecting a total of 1.2M executions of queries for which the result page contained at least one enhanced snippet among the first ten results. Out of those queries we filtered out navigational sessions. Our definition of *navigational sessions* includes query executions that contain one relevant result in the first position, where the user clicked on that result and no further clicks were made. This helps us to clear out queries which are non-informative for the purpose of the experiment (like users issuing the query *twitter* and clicking on *www.twitter.com*). This resulted in 530K query sessions. Query and URL relevance was assesed by trained editors, as in previous experiments. The editors were asked to grade each pair on a 5-point scale (from *Not relevant* to *Perfect*); we consider a result relevant if its grade is higher than 2.

Table 7: Estimation of the different probabilities for the click model

Probability	Text Snippet	Enhanced	Change
$p(R\|E)$	0.320	0.417	+30.3%
$p(R\|V, E)$	0.376	0.468	+24.47%
$p(R\|C, E)$	0.286	0.533	+86.36%
$p(C\|V, E)$	0.051	0.019	-62.745%

Our model shows that on average a user examined 5.8 results per query and clicked on 1.5 results. Table 7 contains the different probabilities presented above. We compare the results of traditional textual snippets (column *text snippets*) with those of enhanced snippets (column *enhanced*). We assume that the random variables R (relevant), V (viewed), and C (click) are set to 1, whereas $E = 0$ for text snippets and $E = 1$ in the enhanced results column. The four rows thus correspond to one of each of the questions posed above.

Our results show that documents that trigger an enhanced snippet in the query sample, $p(R = 1|E = 1)$ are more likely to be relevant *a priori*, than traditional textual snippets $p(R = 1|E = 0)$, which is not surprising given that more enhanced results are generated for head sites than tail sites. The fact that enhanced results are relevant is also noticeable from the resuls displayed to the user: from all the results that the users have actually viewed, there is a 24.5% higher odds that result is relevant if it has an enhanced snippet $p(R = 1|V = 1, E = 1)$.

The most striking result is that even if non-enhanced results get a higher probability of clicking, $p(C = 1|V = 1, E = 0) > p(C = 1|V = 1, E = 1)$ as shown in the last row, the probability that users click on a relevant result, $p(R = 1|C = 1, E = 1)$, is 86.6% higher if the snippet contains an enhanced presentation. Please note that the probability of clicking is computed differently than the CTR of the previous experiment, where we only took into account *good* clicks, whereas this probability reflects all the clicks. In practice, this means that the enhanced results presentation was extremely useful in determining whether the document was relevant or not, and therefore likely to lower the overall search effort, lead to faster task completion and user satisfaction. This is a very positive finding.

5.2 Search result coverage

As alluded to in Section 2, there is a potential that a surprisingly small amount of metadata can power a large proportion of search results. Based on a sampling of the results generated by Yahoo search engine in November 2009, Table 8 shows a relatively small amount of data available. However, Table 9 proves that the majority of search results can contain at least one enhanced result with very little metadata in the crawled corpus.[9]

Table 8: Metadata availability within the Yahoo search corpus

Document type	% of corpus
All documents with metadata from any source	5.8%
All documents with RDFa- or microformat-based metadata	4.0%
All documents with enough metadata providing an enhanced result	2.3%

As called out in the side-by-side tests, improperly extracted key-value pairs have a profound impact on the usefulness of the data provided by enhanced results. To test the data, we exported 1000 URLs matching specific templates and asked internal editorial resources to evaluate the quality

[9]Recall from Section 2.1, Wikipedia can appear in over 25% of the search results

Table 9: Percentage of search result pages showing an enhanced search result

Template	Page views
Reference	28.2%
Product	8.4%
Person	5.9%
Video	5.5%
Local	4.3%
Event	1.1%
Other	0.3%

of extraction. To distinguish the two main sources of incorrect data, we asked evaluators to mark the data extracted data as either "correct" (extracted data matches the cached crawl content) or "bad" (extracted data does not match the cached crawl content).

As shown in Table 10, data source quality varies greatly by content contributor.[10] We found that feeds and general page markup had the greatest variance in data quality, and wrappers, page markup created specifically for SearchMonkey, and human-created extraction rules created for specific websites had the highest precision. During the analysis, we realized that the early results of data quality for generally available RDFa and microformat page markup was so low ($< 40\%$) that we terminated that analysis and excluded those sources from the remainder of the review process.

Table 10: Subset of source quality analysis

Template	Source Type	Source ID	Precision
Product	feed	0	59.2%
Product	human-created rule	1	92.8%
Product	wrapper	2	91.5%
Product	feed	3	86.9%
...	...	...	
Video	human-created	n	81.2%

Overall, the recommendation by the editorial reviewers was to consider the quality of metadata sources in the following order:

- page markup created specifically for our system SearchMonkey
- human-created extraction rule
- supervised wrapper induction
- partner feeds
- general page markup, such as RDFa and microformats

Precision was consistently high for markup created specifically for SearchMonkey. Wrapper-based entity extraction ($> 90\%$) also exhibits high quality on average, but with a larger variance. Some sites are very durable, and others frequently change their DOM structure often as a defense against page scrapers. However, this behavior is traceable and problematic sites can be removed from extraction.

[10]The data in Table 10 has been partly hidden, but we include this table as a representation of the data captured during the experiment.

In terms of reusing existing markup not specifically provided for our tool, we observed that popular microformats are often misused, while emerging, small microformats exhibit large variations in syntax. The hcard microformat is very popular on the Web, appearing in over 500M Web documents. Due to its popularity, hcard is often misused to represent data (such as news) for which its not intended. Due to the lack of extensibility of microformats, users also often abuse existing fields for representing information that could not be captured otherwise. Smaller microformats such as hrecipe and hproduct are problematic due to different versions being in use at the same time. Further, site owners often mark up data that is not representative of the main content of the page, e.g. an address that is within the template used to generate all pages of the site.

We have also found that relevant images improve the user's experience, but irrelevant images can actually lead to a negative experience in which no image would have been a better presentation. Consider various queries about Switzerland with the below presentation:

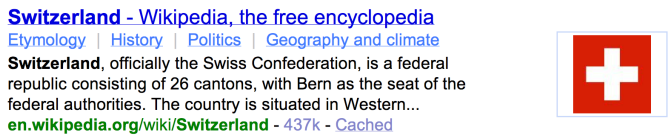

Figure 6: Example of an enhanced search result for "Switzerland"

- "Switzerland": the flag can provide a slightly positive or negative experience. Some users felt that the flag reinforces they are looking at the "right" Switzerland as the image confirmed the search engine found the right intent. Other users felt it was not helpful or possibly used space that could have provided better information from the snippet.
- "Switzerland flag": the flag is exactly what the user is looking for, providing a first result similar to what the user would receive using an image search engine
- "Switzerland map": the flag is irrelevant to the query intent, and opens the possibility that the search engine does not know the difference between a map and a flag

Likewise, the tolerance of adult content is lower for images than text, and can lead to interesting discussions whether visual content has a different threshold than textual content. Wikipedia, generally assumed to not be a source of adult rich media on the Web, does contain topics in which the textual summaries are not offensive by Yahoo editorial standards, but the image itself is offensive (or at least borderline). An example is the article on "pubic hair" which is considered generally acceptable when presented using a textual abstract. However, the article also contains graphic depictions, which are not considered appropriate. This discrepency between textual offensiveness scoring and image offensiveness scoring needs to be considered when generating enhanced results.

6. CONCLUSIONS

We have proposed the notion of enhanced search results as an extension of traditional search result display with images, key-value pairs, and other interactive elements. The generation of enhanced results is enabled by the standards and adoption of the Semantic Web and advances in Information Extraction. Based on this opportunity, we have implemented an extension to our search infrastructure that allows both users and search engine providers to easily define plugins that translate web page metadata into search result displays.

We conducted three major experiments to assess the effectiveness of enhanced results. In the context of a user study, we have shown that users prefer enhanced search results over the traditional search result by a large margin when explicitly asked to make a side-by-side comparison. We have also observed users under their normal search behavior in an online experiment that involved exposing part of our search traffic to an alternative search result page where enhanced results have been enabled. Compared to the activity on the baseline search result pages showing text-only abstracts, we have measured a higher CTR (excluding bad clicks) for enhanced results. This is a key finding because the CTR metric is widely used in search engine optimization, and higher CTRs may convince page owners to implement semantic markup, which in turn allows search engines to display enhanced results and provide other services.

By applying a more expressive user interaction model, we have also shown that enhanced results effectively guide the users to relevant content in general. This is a substantial finding that points to the effectiveness of enhanced results, as the enhancement of the snippet with structured data lowers the likelihood of bad clicks. It also provides ample motivation for publishers of web content to consider enhanced results as a way to avoid "bad clicks" and attract the searchers who would find their content relevant.

We have also shown in our fourth experiment that even though a relatively low number of web pages in our index contains explicit or extracted metadata, a large fraction of the search result pages feature enhanced results, due to the uneven distribution of user attention across content sources. This proves that despite the fragmentation of web content, we can generate enhanced results efficiently, and have a significant impact on the search experience of all users.

7. FUTURE WORK

Although enhanced results as described, implemented and evaluated in this paper are themselves radical departures from how search result abstracts are currently generated, there are a number of potential areas for further development.

As an example, existing studies show that structured data may play an even larger role in search from mobile devices, a growing area of both research and development. Church et. al [5] show that transactional and navigational queries are significantly more frequent in mobile usage compared to web search. Informational queries receive much less weight in mobile scenarios due to the amount of content that users can consume on the mobile: long running searches that require research, possibly using multiple sources, are typically missing. Instead, mobile search provides a larger role in getting information related to particular objects (places, events, schedules) near the user. The same study also shows an overwhelming dominance of adult queries due to the much more private sphere of mobile search. We expect that metadata-based search and result presentation will play a significant role in mobile search in the future.

User experience could be also further improved by allowing additional context to be taken into account in generating web search results. Although our current enhanced results are static, we foresee the possibility to modify the result presentation based on the query, or a structured interpretation of the query. For example, when the user is asking for the location of a restaurant, it makes sense to highlight location information in the result, if necessary by suppressing more generic information. Plug-ins may also take more than one result as input, leading to applications that can help the user aggregate information from multiple result documents. In fact, query aware applications may be triggered by the query itself, showing up dynamically as required.

8. ACKNOWLEDGEMENTS

We would like to acknowledge Amit Kumar, Micah Dubinko, Sara Golemon, Chris Lindsey, Weilin Ng and all other past and present members of the Yahoo! Search team who have helped us to realize SearchMonkey.

9. REFERENCES

[1] B. Adida, M. Birbeck, S. McCarron, and S. Pemberton. RDFa in XHTML: Syntax and Processing, October 2008. http://www.w3.org/TR/2008/REC-rdfa-syntax-20081014.

[2] X. Bai, R. Delbru, and G. Tummarello. RDF Snippets for Semantic Web Search Engines. In *OTM '08: Proceedings of the OTM 2008 Confederated International Conferences, CoopIS, DOA, GADA, IS, and ODBASE 2008. Part II on On the Move to Meaningful Internet Systems*, pages 1304–1318, Berlin, Heidelberg, 2008. Springer-Verlag.

[3] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th international conference on the World Wide Web - WWW '09*, pages 1–10, New York, NY, USA, 2009. ACM.

[4] G. Cheng, W. Ge, and Y. Qu. Falcons: searching and browsing entities on the semantic web. In *Proceedings of the 17th international conference on the World Wide Web - WWW '08*, pages 1101–1102, New York, New York, USA, 2008. ACM Press.

[5] K. Church, B. Smyth, K. Bradley, and P. Cotter. A large scale study of European mobile search behaviour. In *MobileHCI '08: Proceedings of the 10th international conference on Human computer interaction with mobile devices and services*, pages 13–22, New York, NY, USA, 2008. ACM.

[6] C. L. A. Clarke, E. Agichtein, S. Dumais, and R. W. White. The influence of caption features on clickthrough patterns in web search. In *SIGIR '07*, pages 135–142, New York, NY, USA, 2007. ACM.

[7] E. Cutrell and Z. Guan. What are you looking for?: an eye-tracking study of information usage in web search. In *Proceedings of the SIGCHI conference on*

Human factors in computing systems, CHI '07, pages 407–416, New York, NY, USA, 2007. ACM.

[8] M. d'Aquin, C. Baldassarre, L. Gridinoc, S. Angeletou, M. Sabou, and E. Motta. Characterizing knowledge on the semantic web with watson. In *EON*, 2007.

[9] A. L. David Ciemiewicz, Tapas Kanungo and M. Stone. On the use of long dwell time clicks for measuring user satisfaction with application to web summarization. In *Yahoo! Research Technical Report*, 2010.

[10] L. Ding, T. Finin, A. Joshi, R. Pan, S. R. Cost, Y. Peng, P. Reddivari, V. Doshi, and J. Sachs. Swoogle: a search and metadata engine for the semantic web. In *CIKM '04: Proceedings of the thirteenth ACM conference on Information and knowledge management*, pages 652–659, New York, NY, USA, 2004. ACM Press.

[11] S. Dumais, E. Cutrell, and H. Chen. Optimizing search by showing results in context. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '01, pages 277–284, New York, NY, USA, 2001. ACM.

[12] G. Dupret, V. Murdock, and B. Piwowarski. Web search engine evaluation using clickthrough data and a user model. In *Proceedings of the Workshop on Query Log Analysis*, 2007.

[13] G. E. Dupret and B. Piwowarski. A user browsing model to predict search engine click data from past observations. In *SIGIR '08*, page 331, New York, New York, USA, 2008. ACM Press.

[14] M. Fernandez, V. Lopez, M. Sabou, V. Uren, D. Vallet, E. Motta, and P. Castells. Semantic Search Meets the Web. In *IEEE Semantic Computing*, pages 253–260, 2008.

[15] A. Harth, J. Umbrich, A. Hogan, and S. Decker. YARS2: A Federated Repository for Querying Graph Structured Data from the Web. *The Semantic Web*, pages 211–224, 2008.

[16] J. He, C. Zhai, and X. Li. Evaluation of methods for relative comparison of retrieval systems based on clickthroughs. In *Proceedings of the 18th ACM conference on Information and knowledge management - CIKM '09*, page 2029, New York, New York, USA, 2009. ACM Press.

[17] Y. Huang, Z. Liu, and Y. Chen. Query biased snippet generation in XML search. In *SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 315–326, New York, NY, USA, 2008. ACM.

[18] T. Kanungo, N. Ghamrawi, K. Y. Kim, and L. Wai. Web search result summarization: title selection algorithms and user satisfaction. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pages 1581–1584, New York, NY, USA, 2009. ACM.

[19] T. Kanungo and D. Orr. Predicting the readability of short web summaries. In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 202–211, New York, NY, USA, 2009. ACM.

[20] N. Kushmerick, D. S. Weld, and R. B. Doorenbos. Wrapper induction for information extraction. In *IJCAI*, pages 729–737, 1997.

[21] K. McKeown, R. J. Passonneau, D. K. Elson, A. Nenkova, and J. Hirschberg. Do summaries help? In *SIGIR '05*, pages 210–217, New York, NY, USA. ACM.

[22] G. H. McLaughlin. SMOG Grading-a New Readability Formula. *Journal of Reading*, 12(8):639–646, 1969.

[23] P. Mika, E. Meij, and H. Zaragoza. Investigating the semantic gap through query log analysis. In *Proceedings of the International Semantic Web Conference - ISWC '09*, volume 5823 of *Lecture Notes in Computer Science*, pages 441–455. Springer, 2009.

[24] E. Oren, R. Delbru, M. Catasta, R. Cyganiak, H. Stenzhorn, and G. Tummarello. Sindice.com: a document-oriented lookup index for open linked data. *IJMSO*, 3(1):37–52, 2008.

[25] T. Penin, H. Wang, T. Tran, and Y. Yu. Snippet Generation for Semantic Web Search Engines. In *ASWC '08: Proceedings of the 3rd Asian Semantic Web Conference on The Semantic Web*, pages 493–507, Berlin, Heidelberg, 2008. Springer-Verlag.

[26] E. Pietriga, C. Bizer, D. Karger, and R. Lee. Fresnel: A Browser-Independent Presentation Vocabulary for RDF. In *The Semantic Web - ISWC 2006*, pages 158–171, 2006.

[27] E. Prud'hommeaux and A. Seaborne. SPARQL Query Language for RDF, January 2008. http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/.

[28] J.-T. Sun, D. Shen, H.-J. Zeng, Q. Yang, Y. Lu, and Z. Chen. Web-page summarization using clickthrough data. In R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait, editors, *SIGIR'05*, pages 194–201. ACM, 2005.

[29] T. Tran, H. Wang, and P. Haase. SearchWebDB: Data Web Search on a Pay-As-You-Go Integration Infrastructure, 2008.

[30] A. Turpin, Y. Tsegay, D. Hawking, and H. E. Williams. Fast generation of result snippets in web search. In *SIGIR'07*, pages 127–134, 2007.

[31] R. Varadarajan and V. Hristidis. Structure-based query-specific document summarization. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 231–232, New York, NY, USA, 2005. ACM.

[32] H. Wang, P. Haase, R. Studer, T. Penin, K. Xu, J. Chen, X. Sun, L. Fu, Q. Liu, Y. Yu, and T. Tran. Hermes: a travel through semantics on the data web. In *Proceedings of the 35th SIGMOD international conference on Management of data - SIGMOD '09*, page 1135. ACM Press, 2009.

[33] R. W. White, I. Ruthven, and J. M. Jose. Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes. In *SIGIR'02*, pages 57–64, New York, NY, USA, 2002. ACM.

[34] L. Zhang, Q. Liu, J. Zhang, H. Wang, Y. Pan, and Y. Yu. Semplore: An ir approach to scalable hybrid query of semantic web data. In *In Proc. of the 7th Intl. Semantic Web Conference*, pages 652–665, 2008.