The Influence of Caption Features on Clickthrough Patterns in Web Search

Charles L. A. Clarke
University of Waterloo

Eugene Agichtein
Emory University

Susan Dumais and Ryen W. White
Microsoft Research

ABSTRACT

Web search engines present lists of *captions*, comprising title, snippet, and URL, to help users decide which search results to visit. Understanding the influence of features of these captions on Web search behavior may help validate algorithms and guidelines for their improved generation. In this paper we develop a methodology to use clickthrough logs from a commercial search engine to study user behavior when interacting with search result captions. The findings of our study suggest that relatively simple caption features such as the presence of all terms query terms, the readability of the snippet, and the length of the URL shown in the caption, can significantly influence users' Web search behavior.

Categories and Subject Descriptors

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

General Terms

Experimentation, Human Factors

Keywords

Web search, summarization, snippets, query logs

1. INTRODUCTION

The major commercial Web search engines all present their results in much the same way. Each search result is described by a brief *caption*, comprising the URL of the associated Web page, a title, and a brief summary (or "snippet") describing the contents of the page. Often the snippet is extracted from the Web page itself, but it may also be taken from external sources, such as the human-generated summaries found in Web directories.

Figure 1 shows a typical Web search, with captions for the top three results. While the three captions share the same basic structure, their content differs in several respects. The snippet of the third caption is nearly twice as long as that of the first, while the snippet is missing entirely from the second caption. The title of the third caption contains all of the query terms in order, while the titles of the first and second captions contain only two of the three terms. One of the query terms is repeated in the first caption. All of the query terms appear in the URL of the third caption, while none appear in the URL of the first caption. The snippet of the first caption consists of a complete sentence that concisely describes the associated page, while the snippet of the third caption consists of two incomplete sentences that are largely unrelated to the overall contents of the associated page and to the apparent intent of the query.

While these differences may seem minor, they may also have a substantial impact on user behavior. A principal motivation for providing a caption is to assist the user in determining the relevance of the associated page without actually having to click through to the result. In the case of a *navigational* query — particularly when the destination is well known — the URL alone may be sufficient to identify the desired page. But in the case of an *informational* query, the title and snippet may be necessary to guide the user in selecting a page for further study, and she may judge the relevance of a page on the basis of the caption alone.

When this judgment is correct, it can speed the search process by allowing the user to avoid unwanted material. When it fails, the user may waste her time clicking through to an inappropriate result and scanning a page containing little or nothing of interest. Even worse, the user may be misled into skipping a page that contains desired information.

All three of the results in figure 1 are relevant, with some limitations. The first result links to the main Yahoo Kids! homepage, but it is then necessary to follow a link in a menu to find the main page for games. Despite appearances, the second result links to a surprisingly large collection of online games, primarily with environmental themes. The third result might be somewhat disappointing to a user, since it leads to only a single game, hosted at the Centers for Disease Control, that could not reasonably be described as "online". Unfortunately, these page characteristics are not entirely reflected in the captions.

In this paper, we examine the influence of caption features on user's Web search behavior, using clickthroughs extracted from search engines logs as our primary investigative tool. Understanding this influence may help to validate algorithms and guidelines for the improved generation of the

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.

Figure 1: ***Top three results for the query:*** **kids online games.**

captions themselves. In addition, these features can play a role in the process of inferring relevance judgments from user behavior [1]. By better understanding their influence, better judgments may result.

Different caption generation algorithms might select snippets of different lengths from different areas of a page. Snippets may be generated in a *query-independent* fashion, providing a summary of the page as a whole, or in a *query-dependent* fashion, providing a summary of how the page relates to the query terms. The correct choice of snippet may depend on aspects of both the query and the result page. The title may be taken from the HTML header or extracted from the body of the document [8]. For links that re-direct, it may be possible to display alternative URLs. Moreover, for pages listed in human-edited Web directories such as the Open Directory Project[1], it may be possible to display alternative titles and snippets derived from these listings.

When these alternative snippets, titles and URLs are available, the selection of an appropriate combination for display may be guided by their features. A snippet from a Web directory may consist of complete sentences and be less fragmentary than an extracted snippet. A title extracted from the body may provide greater coverage of the query terms. A URL before re-direction may be shorter and provide a clearer idea of the final destination.

The work reported in this paper was undertaken in the context of the Windows Live search engine. The image in figure 1 was captured from Windows Live and cropped to eliminate branding, advertising and navigational elements. The experiments reported in later sections are based on Windows Live query logs, result pages and relevance judgments collected as part of ongoing research into search engine performance [1, 2]. Nonetheless, given the similarity of caption formats across the major Web search engines we believe the results are applicable to these other engines. The query in figure 1 produces results with similar relevance on the other major search engines. This and other queries produce captions that exhibit similar variations. In addition, we believe our methodology may be generalized to other search applications when sufficient clickthrough data is available.

[1] www.dmoz.org

2. RELATED WORK

While commercial Web search engines have followed similar approaches to caption display since their genesis, relatively little research has been published about methods for generating these captions and evaluating their impact on user behavior. Most related research in the area of document summarization has focused on newspaper articles and similar material, rather than Web pages, and has conducted evaluations by comparing automatically generated summaries with manually generated summaries. Most research on the display of Web results has proposed substantial interface changes, rather than addressing details of the existing interfaces.

2.1 Display of Web results

Varadarajan and Hristidis [16] are among the few who have attempted to improve directly upon the snippets generated by commercial search systems, without introducing additional changes to the interface. They generated snippets from spanning trees of document graphs and experimentally compared these snippets against the snippets generated for the same documents by the Google desktop search system and MSN desktop search system. They evaluated their method by asking users to compare snippets from the various sources.

Cutrell and Guan [4] conducted an eye-tracking study to investigate the influence of snippet length on Web search performance and found that the optimal snippet length varied according to the task type, with longer snippets leading to improved performance for informational tasks and shorter snippets for navigational tasks.

Many researchers have explored alternative methods for displaying Web search results. Dumais et al. [5] compared an interface typical of those used by major Web search engines with one that groups results by category, finding that users perform search tasks faster with the category interface. Paek et al. [12] propose an interface based on a fisheye lens, in which mouse hovers and other events cause captions to zoom and snippets to expand with additional text.

White et al. [17] evaluated three alternatives to the standard Web search interface: one that displays expanded summaries on mouse hovers, one that displays a list of top ranking sentences extracted from the results taken as a group, and one that updates this list automatically through *implicit feedback*. They treat the length of time that a user spends viewing a summary as an implicit indicator of relevance. Their goal was to improve the ability of users to interact with a given result set, helping them to look beyond the first page of results and to reduce the burden of query re-formulation.

2.2 Document summarization

Outside the narrow context of Web search considerable related research has been undertaken on the problem of document summarization. The basic idea of extractive summarization — creating a summary by selecting sentences or fragments — goes back to the foundational work of Luhn [11]. Luhn's approach uses term frequencies to identify "significant words" within a document and then selects and extracts sentences that contain significant words in close proximity.

A considerable fraction of later work may be viewed as extending and tuning this basic approach, developing improved methods for identifying significant words and selecting sentences. For example, a recent paper by Sun et al. [14] describes a variant of Luhn's algorithm that uses clickthrough data to identify significant words. At its simplest, snippet generation for Web captions might also be viewed as following this approach, with query terms taking on the role of significant words.

Since 2000, the annual Document Understanding Conference (DUC) series, conducted by the US National Institute of Standards and Technology, has provided a vehicle for evaluating much of the research in document summarization[2]. Each year DUC defines a methodology for one or more experimental tasks, and supplies the necessary test documents, human-created summaries, and automatically extracted baseline summaries. The majority of participating systems use extractive summarization, but a number attempt natural language generation and other approaches.

Evaluation at DUC is achieved through comparison with manually generated summaries. Over the years DUC has included both single-document summarization and multi-document summarization tasks. The main DUC 2007 task is posed as taking place in a question answering context. Given a topic and 25 documents, participants were asked to generate a 250-word summary satisfying the information need enbodied in the topic. We view our approach of evaluating summarization through the analysis of Web logs as complementing the approach taken at DUC.

A number of other researchers have examined the value of query-dependent summarization in a non-Web context. Tombros and Sanderson [15] compared the performance of 20 subjects searching a collection of newspaper articles when guided by query-independent vs. query-dependent snippets. The query-independent snippets were created by extracting the first few sentences of the articles; the query-dependent snippets were created by selecting the highest scoring sentences under a measure biased towards sentences containing query terms. When query-dependent summaries were presented, subjects were better able to identify relevant documents without clicking through to the full text.

Goldstein et al. [6] describe another extractive system for generating query-dependent summaries from newspaper articles. In their system, sentences are ranked by combining statistical and linguistic features. They introduce normalized measures of recall and precision to facilitate evaluation.

2.3 Clickthroughs

Queries and clickthroughs taken from the logs of commercial Web search engines have been widely used to improve the performance of these systems and to better understand how users interact with them. In early work, Broder [3] examined the logs of the AltaVista search engine and identified three broad categories of Web queries: informational, navigational and transactional. Rose and Levinson [13] conducted a similar study, developing a hierarchy of query goals with three top-level categories: informational, navigational and resource. Under their taxonomy, a transactional query as defined by Broder might fall under either of their three categories, depending on details of the desired transaction.

Lee et al. [10] used clickthrough patterns to automatically categorize queries into one of two categories: informational — for which multiple Websites may satisfy all or part of the user's need — and navigational — for which users have a particular Website in mind. Under their taxonomy, a transactional or resource query would be subsumed under one of these two categories.

Agichtein et al. interpreted caption features, clickthroughs and other user behavior as implicit feedback to learn preferences [2] and improve ranking [1] in Web search. Xue et al. [18] present several methods for associating queries with documents by analyzing clickthrough patterns and links between documents. Queries associated with documents in this way are treated as meta-data. In effect, they are added to the document content for indexing and ranking purposes.

Of particular interest to us is the work of Joachims et al. [9] and Granka et al. [7]. They conducted eye-tracking studies and analyzed log data to determine the extent to which clickthrough data may be treated as implicit relevance judgments. They identified a "trust bias", which leads users to prefer the higher ranking result when all other factors are equal. In addition, they explored techniques that treat clicks as pairwise preferences. For example, a click at position $N+1$ — after skipping the result at position N — may be viewed as a preference for the result at position $N+1$ relative to the result at position N. These findings form the basis of the clickthrough inversion methodology we use to interpret user interactions with search results. Our examination of large search logs compliments their detailed analysis of a smaller number of participants.

3. CLICKTHROUGH INVERSIONS

While other researchers have evaluated the display of Web search results through user studies — presenting users with a small number of different techniques and asking them to complete experimental tasks — we approach the problem

[2] `duc.nist.gov`

by extracting implicit feedback from search engine logs. Examining user behavior *in situ* allows us to consider many more queries and caption characteristics, with the volume of available data compensating for the lack of a controlled lab environment.

The problem remains of interpreting the information in these logs as implicit indicators of user preferences, and in this matter we are guided by the work of Joachims et al. [9]. We consider *caption pairs*, which appear adjacent to one another in the result list.

Our primary tool for examining the influence of caption features is a type of pattern observed with respect to these caption pairs, which we call a *clickthrough inversion*. A clickthrough inversion occurs at position N when the result at position N receives fewer clicks than the result at position $N + 1$. Following Joachims et al. [9], we interpret a clickthrough inversion as indicating a preference for the lower ranking result, overcoming any trust bias. For simplicity, in the remainder of this paper we refer to the higher ranking caption in a pair as "caption A" and the lower ranking caption as "caption B".

3.1 Extracting clickthroughs

For the experiments reported in this paper, we sampled a subset of the queries and clickthroughs from the logs of the Windows Live search engine over a period of 3-4 days on three separate occasions: once for results reported in section 3.3, once for a pilot of our main experiment, and once for the experiment itself (sections 4 and 5). For simplicity we restricted our sample to queries submitted to the US English interface and ignored any queries containing complex or non-alphanumeric terms (e.g. operators and phrases). At the end of each sampling period, we downloaded captions for the queries associated with the clickthrough sample.

When identifying clickthroughs in search engine logs, we consider only the first clickthrough action taken by a user after entering a query and viewing the result page. Users are identified by IP address, which is a reasonably reliable method of eliminating multiple results from a single user, at the cost of falsely eliminating results from multiple users sharing the same address.

By focusing on the initial clickthrough, we hope to capture a user's impression of the relative relevance within a caption pair when first encountered. If the user later clicks on other results or re-issues the same query, we ignore these actions. Any preference captured by a clickthrough inversion is therefore a preference among a group of users issuing a particular query, rather than a preference on the part of a single user. In the remainder of the paper, we use the term "clickthrough" to refer only to this initial action.

Given the dynamic nature of the Web and the volumes of data involved, search engine logs are bound to contain considerable "noise". For example, even over a period of hours or minutes the order of results for a given query can change, with some results dropping out of the top ten and new ones appearing. For this reason, we retained clickthroughs for a specific combination of a query and a result only if this result appears in a consistent position for at least 50% of the clickthroughs. Clickthroughs for the same result when it appeared at other positions were discarded. For similar reasons, if we did not detect at least ten clickthroughs for a particular query during the sampling period, no clickthroughs for that query were retained.

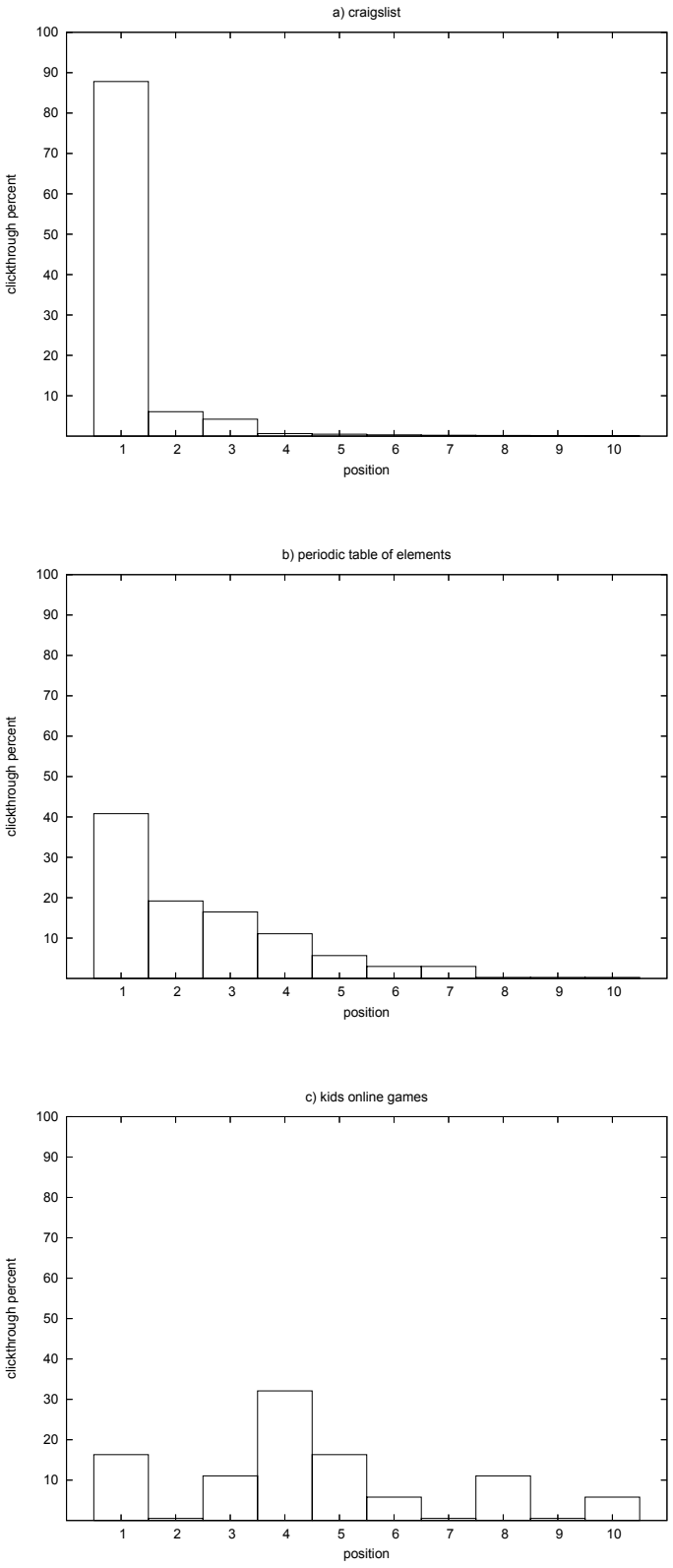

Figure 2: ***Clickthrough curves for three queries: a) a stereotypical navigational query, b) a stereotypical informational query, and c) a query exhibiting clickthrough inversions.***

The outcome at the end of each sampling period is a set of records, with each record describing the clickthroughs for a given query/result combination. Each record includes a query, a result position, a title, a snippet, a URL, the number of clickthroughs for this result, and the total number of clickthroughs for this query. We then processed this set to generate clickthrough curves and identify inversions.

3.2 Clickthrough curves

It could be argued that under ideal circumstances, clickthrough inversions would not be present in search engine logs. A hypothetical "perfect" search engine would respond to a query by placing the result most likely to be relevant first in the result list. Each caption would appropriately summarize the content of the linked page and its relationship to the query, allowing users to make accurate judgments. Later results would complement earlier ones, linking to novel or supplementary material, and ordered by their interest to the greatest number of users.

Figure 2 provides clickthrough curves for three example queries. For each example, we plot the percentage of clickthroughs against position for the top ten results. The first query (`craigslist`) is stereotypically navigational, showing a spike at the "correct" answer (`www.craigslist.org`). The second query is informational in the sense of Lee et al. [10] (`periodic table of elements`). Its curve is flatter and less skewed toward a single result. For both queries, the number of clickthroughs is consistent with the result positions, with the percentage of clickthroughs decreasing monotonically as position increases, the ideal behavior.

Regrettably, no search engine is perfect, and clickthrough inversions are seen for many queries. For example, for the third query (`kids online games`) the clickthrough curve exhibits a number of clickthrough inversions, with an apparent preference for the result at position 4.

Several causes may be enlisted to explain the presence of an inversion in a clickthrough curve. The search engine may have failed in its primary goal, ranking more relevant results below less relevant results. Even when the relative ranking is appropriate, a caption may fail to reflect the content of the underlying page with respect to the query, leading the user to make an incorrect judgment. Before turning to the second case, we address the first, and examine the extent to which relevance alone may explain these inversions.

3.3 Relevance

The simplest explanation for the presence of a clickthrough inversion is a relevance difference between the higher ranking member of caption pair and the lower ranking member. In order to examine the extent to which relevance plays a role in clickthrough inversions, we conducted an initial experiment using a set of 1,811 queries with associated judgments created as part of on-going work. Over a four-day period, we sampled the search engine logs and extracted over one hundred thousand clicks involving these queries. From these clicks we identified 355 clickthrough inversions, satisfying the criteria of section 3.1, where relevance judgments existed for both pages.

The relevance judgments were made by independent assessors viewing the pages themselves, rather than the captions. Relevance was assessed on a 6-point scale. The outcome is presented in figure 3, which shows the explicit judgments for the 355 clickthrough inversions. In all of these cases, there were more clicks on the lower ranked member of the pair (B). The figure shows the corresponding relevance judgments. For example, the first row rel(A) < rel(B), indicates that the higher ranking member of pair (A) was rated as less relevant than the lower ranking member of the pair (B).

Relationship	Number	Percent
rel(A) < rel(B)	119	33.5%
rel(A) = rel(B)	134	37.7%
rel(A) > rel(B)	102	28.7%

Figure 3: ***Relevance relationships at clickthrough inversions. Compares relevance between the higher ranking member of a caption pair*** **(rel(A))** ***to the relevance of the lower ranking member*** **(rel(B))***, where caption A received fewer clicks than caption B.***

As we see in the figure, relevance alone appears inadequate to explain the majority of clickthrough inversions. For two-thirds of the inversions (236), the page associated with caption A is at least as relevant as the page associated with caption B. For 28.7% of the inversions, A has greater relevance than B, which received the greater number of clickthroughs.

4. INFLUENCE OF CAPTION FEATURES

Having demonstrated that clickthrough inversions cannot always be explained by relevance differences, we explore what features of caption pairs, if any, lead users to prefer one caption over another. For example, we may hypothesize that the absence of a snippet in caption A and the presence of a snippet in caption B (e.g. captions 2 and 3 in figure 1) leads users to prefer caption A. Nonetheless, due to competing factors, a large set of clickthrough inversions may also include pairs where the snippet is missing in caption B and not in caption A. However, if we compare a large set of clickthrough inversions to a similar set of pairs for which the clickthroughs are consistent with their ranking, we would expect to see relatively more pairs where the snippet was missing in caption A.

4.1 Evaluation methodology

Following this line of reasoning, we extracted two sets of caption pairs from search logs over a three day period. The first is a set of nearly five thousand clickthrough inversions, extracted according to the procedure described in section 3.1. The second is a corresponding set of caption pairs that do not exhibit clickthrough inversions. In other words, for pairs in this set, the result at the higher rank (caption A) received more clickthroughs than the result at the lower rank (caption B). To the greatest extent possible, each pair in the second set was selected to correspond to a pair in the first set, in terms of result position and number of clicks on each result. We refer to the first set, containing clickthrough inversions, as the INV set; we refer to the second set, containing caption pairs for which the clickthroughs are consistent with their rank order, as the CON set.

We extract a number of features characterizing snippets (described in detail in the next section) and compare the presence of each feature in the INV and CON sets. We describe the features as a hypothesized preference (e.g., a preference for captions containing a snippet). Thus, in either set, a given feature may be present in one of two forms: favoring the higher ranked caption (caption A) or favoring the lower ranked caption (caption B). For example, the ab-

Feature Tag	Description
MissingSnippet	snippet missing in caption A and present in caption B
SnippetShort	short snippet in caption A (< 25 characters) with long snippet (> 100 characters) in caption B
TermMatchTitle	title of caption A contains matches to fewer query terms than the title of caption B
TermMatchTS	title+snippet of caption A contains matches to fewer query terms than the title+snippet of caption B
TermMatchTSU	title+snippet+URL of caption A contains matches to fewer query terms than caption B
TitleStartQuery	title of caption B (but not A) starts with a phrase match to the query
QueryPhraseMatch	title+snippet+url contains the query as a phrase match
MatchAll	caption B contains one match to each term; caption A contains more matches with missing terms
URLQuery	caption B URL is of the form `www.query.com` where the query matches exactly with spaces removed
URLSlashes	caption A URL contains more slashes (i.e. a longer path length) than the caption B URL
URLLenDIff	caption A URL is longer than the caption B URL
Official	title or snippet of caption B (but not A) contains the term "official" (with stemming)
Home	title or snippet of caption B (but not A) contains the phrase "home page"
Image	title or snippet of caption B (but not A) contains a term suggesting the presence of an image gallery
Readable	caption B (but not A) passes a simple readability test

Figure 4: ***Features measured in caption pairs (caption A and caption B), with caption A as the higher ranked result. These features are expressed from the perspective of the prevalent relationship predicted for clickthrough inversions.***

sence of a snippet in caption A favors caption B, and the absence of a snippet in caption B favors caption A. When the feature favors caption B (consistent with a clickthrough inversion) we refer to the caption pair as a "positive" pair. When the feature favors caption A, we refer to it as a "negative" pair. For missing snippets, a positive pair has the caption missing in caption A (but not B) and a negative pair has the caption missing in B (but not A).

Thus, for a specific feature, we can construct four subsets: 1) INV+, the set of positive pairs from INV; 2) INV−, the set of negative pairs from INV; 3) CON+; the set of positive pairs from CON; and 4) CON− the set of negative pairs from CON. The sets INV+, INV−, CON+, and CON− will contain different subsets of INV and CON for each feature. When stating a feature corresponding to a hypothesized user preference, we follow the practice of stating the feature with the expectation that the size of INV+ relative to the size of INV− should be greater than the size of CON+ relative to the size of CON−. For example, we state the missing snippet feature as "snippet missing in caption A and present in caption B".

This evaluation methodology allows us to construct a contingency table for each feature, with INV essentially forming the experimental group and CON the control group. We can then apply Pearson's chi-square test for significance.

4.2 Features

Figure 4 lists the features tested. Many of the features on this list correspond to our own assumptions regarding the importance of certain caption characteristics: the presence of query terms, the inclusion of a snippet, and the importance of query term matches in the title. Other features suggested themselves during the examination of the snippets collected as part of the study described in section 3.3 and during a pilot of the evaluation methodology (section 4.1). For this pilot we collected INV and CON sets of similar sizes, and used these sets to evaluate a preliminary list of features and to establish appropriate parameters for the SnippetShort and Readable features. In the pilot, all of the features list in figure 4 were significant at the 95% level. A small number of other features were dropped after the pilot.

These features all capture simple aspects of the captions. The first feature concerns the existence of a snippet and the second concerns the relative size of snippets. Apart from this first feature, we ignore pairs where one caption has a missing snippet. These pairs are not included in the sets constructed for the remaining features, since captions with missing snippets do not contain all the elements of a standard caption and we wanted to avoid their influence.

The next six features concern the location and number of matching query terms. For the first five, a match for each query term is counted only once, additional matches for the same term are ignored. The MatchAll feature tests the idea that matching all the query terms exactly once is preferable to matching a subset of the terms many times with a least one query term unmatched.

The next three features concern the URLs, capturing aspects of their length and complexity, and the last four features concern caption content. The first two of these content features (Official and Home) suggest claims about the importance or significance of the associated page. The third content feature (Image) suggests the presence of an image gallery, a popular genre of Web page. Terms represented by this feature include "pictures", "pics", and "gallery".

The last content feature (Readable) applies an *ad hoc* readability metric to each snippet. Regular users of Web search engines may notice occasional snippets that consist of little more than lists of words and phrases, rather than a coherent description. We define our own metric, since the Flesch-Kincaid readability score and similar measures are intended for entire documents not text fragments. While the metric has not been experimentally validated, it does reflect our intuitions and observations regarding result snippets. In English, the 100 most frequent words represent about 48% of text, and we would expect readable prose, as opposed to a disjointed list of words, to contain these words in roughly this proportion. The Readable feature computes the percentage of these top-100 words appearing in each caption. If these words represent more than 40% of one caption and less than 10% of the other, the pair is included in the appropriate set.

Feature Tag	INV+	INV−	%+	CON+	CON−	%+	χ^2	p-value
MissingSnippet	185	121	60.4	144	133	51.9	4.2443	0.0393
SnippetShort	20	6	76.9	12	16	42.8	6.4803	0.0109
TermMatchTitle	800	559	58.8	660	700	48.5	29.2154	<.0001
TermMatchTS	310	213	59.2	269	216	55.4	1.4938	0.2216
TermMatchTSU	236	138	63.1	189	149	55.9	3.8088	0.0509
TitleStartQuery	1058	933	53.1	916	1096	45.5	23.1999	<.0001
QueryPhraseMatch	465	346	57.3	427	422	50.2	8.2741	0.0040
MatchAll	8	2	80.0	1	4	20.0		*0.0470*
URLQuery	277	188	59.5	159	315	33.5	63.9210	<.0001
URLSlashes	1715	1388	55.2	1380	1758	43.9	79.5819	<.0001
URLLenDiff	2288	2233	50.6	2062	2649	43.7	43.2974	<.0001
Official	215	142	60.2	133	215	38.2	34.1397	<.0001
Home	62	49	55.8	64	82	43.8	3.6458	0.0562
Image	391	270	59.1	315	335	48.4	15.0735	<.0001
Readable	52	43	54.7	31	48	39.2	4.1518	0.0415

Figure 5: ***Results corresponding to the features listed in figure 4 with χ^2 and p-values ($df = 1$). Features supported at the 95% confidence level are bolded. The p-value for the MatchAll feature is computed using Fisher's Exact Test.***

4.3 Results

Figure 5 presents the results. Each row lists the size of the four sets (INV+, INV−, CON+, and CON−) for a given feature and indicates the percentage of positive pairs (%+) for INV and CON. In order to reject the null hypothesis, this percentage should be significantly greater for INV than CON. Except in one case, we applied the chi-squared test of independence to these sizes, with p-values shown in the last column. For the MatchAll feature, where the sum of the set sizes is 15, we applied Fisher's exact test. Features supported at the 95% confidence level are **bolded**.

5. COMMENTARY

The results support claims that missing snippets, short snippets, missing query terms and complex URLs negatively impact clickthroughs. While this outcome may not be surprising, we are aware of no other work that can provide support for claims of this type in the context of a commercial Web search engine.

This work was originally motivated by our desire to validate some simple guidelines for the generation of captions — summarizing opinions that we formulated while working on related issues. While our results do not direct address all of the many variables that influence users understanding of captions, they are consistent with the major guidelines. Further work is needed to provide additional support for the guidelines and to understand the relationships among variables.

The first of these guidelines underscores the importance of displaying query terms in context: *Whenever possible all of the query terms should appear in the caption, reflecting their relationship to the associated page.* If a query term is missing from a caption, the user may have no idea why the result was returned. The results for the MatchAll feature directly support this guideline. The results for TermMatchTitle and TermMatchTSU confirm that matching more terms is desirable. Other features provide additional indirect support for this guideline, and none of the results are inconsistent with it.

A second guideline speaks to the desirability of presenting the user with a readable snippet: *When query terms are present in the title, they need not be repeated in the snippet.* In particular, when a high-quality query-independent summary is available from an external source, such as a Web directory, it may be more appropriate to display this summary than a lower-quality query-dependent fragment selected on-the-fly. When titles are available from multiple sources —the header, the body, Web directories — a caption generation algorithm might a select a combination of title, snippet and URL that includes as many of the query terms as possible. When a title containing all query terms can be found, the algorithm might select a query-independent snippet. The MatchAll and Readable features directly support this guideline. Once again, other features provide indirect support, and none of the results are inconsistent with it.

Finally, the length and complexity of a URL influences user behavior. When query terms appear in the URL they should highlighted or otherwise distinguished. When multiple URLs reference the same page (due to re-directions, etc.) the shortest URL should be preferred, provided that all query terms will still appear in the caption. In other words, *URLs should be selected and displayed in a manner that emphasizes their relationship to the query.* The three URL features, as well as TermMatchTSU, directly support this guideline.

The influence of the Official and Image features led us to wonder what other terms are prevalent in the captions of clickthrough inversions. As an additional experiment, we treated each of the terms appearing in the INV and CON sets as a separate feature (case normalized), ranking them by their χ^2 values. The results are presented in figure 6. Since we use the χ^2 statistic as a divergence measure, rather than a significance test, no p-values are given. The final column of the table indicates the direction of the influence, whether the presence of the terms positively or negatively influence clickthroughs.

The positive influence of "official" has already been observed (the difference in the χ^2 value from that of figure 5 is due to stemming). None of the terms included in the Image

Rank	Term	χ^2	influence
1	encyclopedia	114.6891	↓
2	wikipedia	94.0033	↓
3	official	36.5566	↑
4	and	28.3349	↑
5	tourism	25.2003	↑
6	attractions	24.7283	↑
7	free	23.6529	↓
8	sexy	21.9773	↑
9	medlineplus	19.9726	↓
10	information	19.9115	↑

Figure 6: ***Words exhibiting the greatest positive (↑) and negative (↓) influence on clickthrough patterns.***

feature appear in the top ten, but "pictures" and "photos" appear at positions 21 and 22. The high rank given to "and" may be related to readability (the term "the" appears in position 20).

Most surprising to us is the negative influence of the terms: "encyclopedia", "wikipedia", "free", and "medlineplus". The first three terms appear in the title of Wikipedia articles[3] and the last appears in the title of MedlinePlus articles[4]. These individual word-level features provide hints about issues. More detailed analyses and further experiments will be required to understand these features.

6. CONCLUSIONS

Clickthrough inversions form an appropriate tool for assessing the influence of caption features. Using clickthrough inversions, we have demonstrated that relatively simple caption features can significantly influence user behavior. To our knowledge, this is first methodology validated for assessing the quality of Web captions through implicit feedback. In the future, we hope to substantially expand this work, considering more features over larger datasets. We also hope to directly address the goal of predicting relevance from clickthoughs and other information present in search engine logs.

7. ACKNOWLEDGMENTS

This work was conducted while the first author was visiting Microsoft Research. The authors thank members of the Windows Live team for their comments and assistance, particularly Girish Kumar, Luke DeLorme, Rohit Wad and Ramez Naam.

8. REFERENCES

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *29th ACM SIGIR*, pages 19–26, Seattle, August 2006.

[2] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting Web search result preferences. In *29th ACM SIGIR*, pages 3–10, Seattle, August 2006.

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

[3] `www.wikipedia.org`

[4] `www.nlm.nih.gov/medlineplus/`

[4] E. Cutrell and Z. Guan. What are you looking for? An eye-tracking study of information usage in Web search. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 407–416, San Jose, California, April-May 2007.

[5] S. Dumais, E. Cutrell, and H. Chen. Optimizing search by showing results in context. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 277–284, Seattle, March-April 2001.

[6] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing text documents: Sentence selection and evaluation metrics. In *22nd ACM SIGIR*, pages 121–128, Berkeley, August 1999.

[7] L. A. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in WWW search. In *27th ACM SIGIR*, pages 478–479, Sheffield, July 2004.

[8] Y. Hu, G. Xin, R. Song, G. Hu, S. Shi, Y. Cao, and H. Li. Title extraction from bodies of HTML documents and its application to Web page retrieval. In *28th ACM SIGIR*, pages 250–257, Salvador, Brazil, August 2005.

[9] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *28th ACM SIGIR*, pages 154–161, Salvador, Brazil, August 2005.

[10] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in Web search. In *14th International World Wide Web Conference*, pages 391–400, Edinburgh, May 2005.

[11] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, April 1958.

[12] T. Paek, S. Dumais, and R. Logan. WaveLens: A new view onto Internet search results. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 727–734, Vienna, Austria, April 2004.

[13] D. Rose and D. Levinson. Understanding user goals in Web search. In *13th International World Wide Web Conference*, pages 13–19, New York, May 2004.

[14] J.-T. Sun, D. Shen, H.-J. Zeng, Q. Yang, Y. Lu, and Z. Chen. Web-page summarization using clickthrough data. In *28th ACM SIGIR*, pages 194–201, Salvador, Brazil, August 2005.

[15] A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In *21st ACM SIGIR*, pages 2–10, Melbourne, Australia, August 1998.

[16] R. Varadarajan and V. Hristidis. A system for query-specific document summarization. In *15th ACM international conference on Information and knowledge management (CIKM)*, pages 622–631, Arlington, Virginia, November 2006.

[17] R. W. White, I. Ruthven, and J. M. Jose. Finding relevant documents using top ranking sentences: An evaluation of two alternative schemes. In *25th ACM SIGIR*, pages 57–64, Tampere, Finland, August 2002.

[18] G.-R. Xue, H.-J. Zeng, Z. Chen, Y. Yu, W.-Y. Ma, W. Xi, and W. Fan. Optimizing web search using Web click-through data. In *13th ACM Conference on Information and Knowledge Management (CIKM)*, pages 118–126, Washington, DC, November 2004.