GENERAL GUIDELINES
ON RANDOM-QUERY EVALUATION

Version 3.1
Last update: December 31, 2003

Random-Query Evaluation

Thank you for participating in one of Google's routine quality control processes, the 'random-query' evaluation. This form of search evaluation takes its name from the fact that the queries which it draws on were randomly selected from our query logs – in other words, these are all queries that someone, at some point, actually entered into the Google query box. Because we want to obtain a realistic impression of how well we're serving the average user, we are careful *not* to pick queries that are particularly well phrased, or easy to search, or unambiguous in intent. (At present we only filter out queries that are clearly pornographic, queries that are complete or incomplete URL addresses such as sport.com , www.simslots.com, ssa.gov or www.California.com and certain numerical queries) That means that you will encounter queries posed by school-age children and reference librarians, research scientists and housekeepers, first-time Internet users and experienced computer geeks. Of course, our decision to include the full spectrum of queries people pose means that evaluation of search results is a tricky business at times, and that – in the absence of help from the person who originally posed the query – we are often confronted with uncertainty about the meaning or purpose of a query and the suitability of the results it brings up.

Let us note from the outset that we evaluate results based on relevance not to a specific person who actually posed the query, but to an imaginary rational mind "behind" the query. Oftentimes, a query may have more than one meaning, or interpretation. In such cases we will have to look at the hypothetical set of rational search engine users behind an ambiguous query, and deduce, or roughly estimate, the make-up of that set; for instance, we will consider the relative presence of zoology enthusiasts and car shoppers in a hypothetical representative sample of the users who could have queried [jaguar].[1]

People use the web, and search in particular, for all sorts of needs and in all sorts of ways. The suitability of results to their searches can be assessed from several perspectives and often along several dimensions. One result on {product ABC} is good if you want to buy the product, but has no information on what to do if ABC malfunctions; another result has the troubleshooting guide for ABC, while a third one does a good job comparing ABC to similar products. There is a certain subjective element involved in evaluation. Despite this complexity, there are some general principles on how to rate query results appropriately and consistently. This document tries to articulate these principles as clearly as possible. Undoubtedly there are many specific situations that it does not cover. If you find yourself repeatedly stumped by a certain type of query and/or result, please do not hesitate to contact your project manager for advice.

---

[1] Throughout this document, we will us square brackets to denote a query exactly as posed to the search engine, including any syntax, e.g. [philosophy+mind], [car dealers, "Mountain View"]. We will use curly brackets to denote a query, or part of a query, by type: {celebrity by name} means a query for anyone who's currently considered a celebrity, [waiters-on-wheels {location}] can mean [waiters-on-wheels san francisco] or [waiters-on-wheels san jose], {global company, location} can stand for [ikea germany] or [hp palo alto], etc.

Introduction

Evaluation process always starts with web coverage research for the query. The major goals of the web coverage research are:

- Determining whether the query is ambiguous, possessing multiple interpretations
- Assigning rough probability values to each interpretation based on available evidence. For example, the query [wallpaper] is ambiguous between the real, tangible wallpaper and the computer wallpaper. Statistically speaking, seeking to download computer wallpaper may be a more likely scenario, but the home improvement interpretation also has its niche on the web. Take a different query, [beer wallpaper], and the desktop decoration interpretation clearly wins over the home improvement one.
- Ascertaining how much, or little, information there is on the web for the query. [2]The knowledge of the query coverage will come in handy when you have to decide whether a particular result merits a relatively high position in the search result listing. For instance, if the query is [Maltese music], a top-level page on a site devoted to regional types of music that has a link to a site on Maltese music among many other links, should not belong in the top ten or so results. Maltese music enjoys good web coverage, and there is no reason to promote a result that's not exactly right.
- Determining the "country of origin" for the query. The default assumption is that the query comes from the US. Broad, "global", results, and specific US results are thus appropriate. But, many queries may override the default assumption. For instance, the query [Motorized bicycle Singapore petrol] uses a word that's not part of American English and specifies a region outside of the United States. Whereas Singapore results to other queries – those that by default fall under the US origin rule – may be inappropriate, Singapore and only Singapore results are appropriate, and may belong in the top ten, for this particular query.

The Query Types

While there is no simple way to categorize all searches into a neatly organized system, three major categories have been used by analysts of web search to draw a distinction between navigational queries, informational queries, and transactional queries. This

[2] Some queries have ample coverage – results to them should be put to the strictest scrutiny. Other queries have scant coverage – rate results to those more leniently. For example, the query ["new mexico" state penal code] cannot bring the desired result, because the State of New Mexico, as of this writing, does not have a state penal code (New Mexico relies instead on its common law).

classification, as it turns out, allows for some useful generalizations in the context of query-result evaluation.

A *navigational* query is one that normally has only one satisfactory result: the user types in the name of an entity (“United Airlines”) and expects to be taken to the homepage of that entity.

An *informational* query can have many or few appropriate results, with varying degrees of relevance/informativeness/utility and varying degrees of authority. The user types in a topic (“renaissance paintings”, “aging disease”), sometimes in the form of an actual question (“What is a quark?”, "How do I...?"), and expects to be provided with information on this topic.

A *transactional* query as well can have many or few appropriate results, of varying quality. However, in this case the user is not requesting information – or at least not only information – but instead has the primary goal of carrying out a transaction. Typically, the transaction consists of the acquisition – for money or free - of a product or service. Some transactions can be fully carried out on the web (think furniture clipart download), some come to fruition offline (think furniture to put in a house).

Again, not every query can be clearly classified. Since products include information products, the line between informational and transactional queries is sometimes hard to draw. Similarly, because the ulterior motive for a navigational query often is to locate a site for potential transactions, there is a grey zone between navigational and transactional queries. To the extent the classification is helpful, use it, but do not attempt to fit any query that comes your way into one of the three boxes: always trying to decide in favor of one or another will only lead to frustration. It may be more helpful to think of different aspects of a query: for instance, the query [thomas the tank engine] can have (a) a navigational aspect - take me to Thomas' homepage, (b) an informational aspect - tell me the history of Thomas creation, and finally, (c) a transactional aspect - I want to buy a book or a toy engine from the Thomas collection.

The Rating Categories

We make use of the following categories for evaluating our search results:

Vital
Useful
Relevant
Off Topic
Offensive

Erroneous
Didn’t Load

Foreign Language
Unrated

Please note that the technical term Relevant differs from the generic word relevant. The same may be true of other category names; to avoid confusion, we always capitalize our category names to set them apart from the generic meanings.

Options from Vital down to Offensive constitute the merit scale. A major subset of the merit scale is the utility continuum, spanning categories from Useful to Off Topic,. Assigning ratings on the merit scale reflects your opinion on where, roughly, the rated results should or could appear in an idealized search result ranking for a given query. Due to the multi-dimensional relativity of the merit scale ratings, more than one rating can be justified at times. In such cases, we ask you to pick the *lower* rating on the scale.

Erroneous, Didn't Load, Foreign Language and Unrated are special categories that are, in effect, non-ratings. By selecting one of these categories, you do not express your opinion on the range of positions the result may occupy in a ranking; rather, you depict certain technical attributes of the result page.

To match the workflow of query-result evaluation, we will start with briefly introducing the non-ratings.[3] For detailed examples, please view the FAQ posted on the Rater Hub.

Didn't Load
A result that's not visible cannot be evaluated. If you are seeing a "Page not found" message, assign Didn't Load. Note that many sites experience a certain amount of downtime on a regular basis. As a result, a page that does not load in the morning may load later in the day or the next day. We would appreciate if you note which results you marked Didn't Loads as you work your way through your project, and briefly revisit them before you sign off on your completed job. Doing so may yield a few extra informative ratings. The in-depth discussion of rating policies in the absence of a working cache is contained in Using Quest, an instruction on the rating interface.

Foreign Language
A result that loads fine but is fully in a foreign language should be labeled as such. For rating English projects, any result that's not in English is Foreign Language (for projects in other languages, please read Appendix "Evaluating i18n results" for the description of the Foreign Language category).[4] Certain exceptions apply when the foreign language page is essentially non-verbal (think images, downloads), and in a few other cases discussed in the answers to FAQ on the Rater Hub.

Erroneous

[3] Please read "Using Quest" to familiarize yourself with the logistics of ratings. "Using Quest" is available online at http://eval.google.com/happier/portal_files/Using_Quest.pdf
[4] The Appendix is NOT required reading if your project is English.

Erroneous results load fine and are not in a foreign language. This category designates what you might think of as "indirect results": an output of searching on an engine or a directory, or a page that offers you to search an engine or directory. Engines and directories that fall in this category search the whole web and not just a subset of it, such as everything in one city or all travel-related information. Of course, Erroneous rating does not apply to engines or directories that are expressly requested by the query.[5]

Unrated
Under certain circumstances, you may be unable to assign a valid rating. For example, despite your best efforts at researching the query and/or result, you may feel you lack certain knowledge to express an opinion. Choose Unrated then.

A page can well possess several of the above technical attributes. We have more discussion on this in the FAQ section of the Rater Hub.

If none of the above technical categories applies, meaning that:

- the page loads fine; and
- the page is in the "correct" language; and
- the page is *not* a search output from an engine or directory (be it Google or another engine/directory organizing information from the web as a whole);[6] and
- you have sufficient information and understanding about the query and/or result,

the result should be rated on the merit scale. We will now introduce the merit scale categories from the top down: from Vital to Offensive.

VERY IMPORTANT: Merit scale evaluation is *not* based on the absence or presence of queried terms on the result page.
Consider a result to the query [German educational toys], http://www.toy-spectrum.com/overview/puzzle/puzzle.html.
Does the absence of the word "educational" reduce the quality of the match in any way? No. The products on the page are clearly educational without being overtly described with this term.[7] Similarly, the query [Users of the internet (Graph)] can be quite adequately answered by a resource that *gives* a graph and mentions such terms as "statistics of", "demographic", "access", "usage", "database", "table" etc., without explicitly mentioning "users" and "Graph".

[5] See [ask jeeves] example in Table 1.
[6] Infrequently, you may see a result that is a Google directory listing, or a result page from another Google interface, or even a search result listing from Google.com. Those instances should all be categorized as Erroneous.
[7] Arguably, a self-respecting educational toy outlet would not mention the word "educational" very much, assuming that customers can recognize a quality product on their own.

VITAL

Vital is a category reserved for very special, unique results to a special subset of queries. Examples illustrate the special attributes of the queries that can have Vital results. Listed in Table 1 are some examples of queries with Vital results to them. You will notice that the queries in Table 1 are predominantly navigational.

Table 1.

Query	**Vital Results**
[OfficeMax]	http://www.officemax.com/
[arizona lottery]	http://www.arizonalottery.com/
[san jose international airport]	http://www.sjc.org/
[Simon and Garfunkel]	http://www.simonandgarfunkel.com/
[suny Binghamton]	http://www.binghamton.edu/
[Union Bank of California]	http://www.uboc.com/uboc/home
[The weather channel]	http://www.weather.com/
[banana republic]	http://www.bananarepublic.com/default.htm
[ask jeeves]	http://www.ask.com/
[interact australia]	http://www.interactaust.com.au/intaust.htm
[los altos school district boundaries map]	http://www.losaltos.k12.ca.us/schls_boundmap.htm#top or http://www.losaltos.k12.ca.us/PDF_Files/Boundaries_2003.pdf (both fit the bill)
[san jose public library]	http://www.sjlibrary.org/
[san jose public library branches]	http://www.sjlibrary.org/about/contacts/branches.htm or http://www.sjlibrary.org/about/locations/index.htm
N-400 form	http://uscis.gov/graphics/formsfee/forms/n-400.htm
[Canadian parliament]	http://www.parl.gc.ca/ or http://www.parl.gc.ca/common/index.asp?Language=E&Parl=37&Ses=2
[disable javascript ie]	http://support.microsoft.com/default.aspx?scid=kb;EN-US;244233 Note that the information may not be displayed conspicuously (in the case at hand, one needs to scroll down the page to read the how-to). The page is not necessarily wholly on the topic of the query. Yet, it provides the how-to endorsed by the creator of IE. Hence, Vital.
[Barbie]	www.barbie.com, from Mattel, the company owning the rights to the brand

Table 1 (cont'd).

Query	Vital Results
[form iap-66]	http://travel.state.gov/visa%3Bexchange.html a page on the site of the ultimate authority on the subject, advises of the change in the title of the form, and therefore appears Vital to the query. Note how Vital results are not necessarily the most useful - but are uniquely authoritative. The pertinent paragraph is buried in the dense text on this page. Someone's personal page shouting "Hey, I searched for IAP-66 and could not find it, guess what, the world has changed and I want everyone to know!" could have been user-friendlier, yet lacking authority.

Is there a Vital result out there for any query imaginable? Emphatically, no. Indeed, *most* queries cannot have Vital results. We will call those queries generic. Generic queries *cannot* have Vital results because no one has ultimate authority on the subject matter and no entity is the uniquely definitive target of the search. .Some queries, such as [things different cultures do for fun] , are obviously generic – no ultimate omniscient authority can ever put together "the" resource for such queries. Other generic queries are sometimes matched by results that may, incorrectly, appear uniquely appropriate. Table 2 lists a few cases in point.

Table 2.

Query	Result that may wrongly appear Vital	Why there are no Vital results
[Learn How To knit]	http://www.learnhowtoknit.com	Please refer to the discussion below on URLs that match the query verbatim.
[crime and punishment]	http://www.nbc.com/Crime_&_Punishment/	Several interpretations "compete" for this query. There is a book by Fyodor Dostoevsky that is part of the global canon and thus an appropriate result for a US-based query. A seminal work by Lawrence Meir Friedman,"Crime and Punishment in American History", is very well known and widely cites in the US. Although it does not match the query fully, the book is commonly referred to as "Crime and Punishment" by Friedman, omitting the rest of the title. Then there is a book "Crime and Punishment in America" by Elliott Currie, another possible interpretation for a U.S.-based query. A handful of resources on law enforcement juxtapose "crime" and "punishment" in their descriptions, evidencing that the word combination has a generic sense to it.
[map of central America]	http://www.infoplease.com/atlas/centralamerica.html	There are terrific resources for maps, but none can claim ultimate authority on the subject of maps of any region

Table 2 (cont'd).

Query	**Result that may wrongly appear Vital**	**Why there are no Vital results**
[mouth ulcers]	(a) http://www.mouthulcers.org/ (b) http://www.nlm.nih.gov/medlineplus/ency/article/001448.htm	Diseases cannot have homepages; no one can claim unique authority to everything related to diagnosing, treating, and preventing any particular disease. Neither a personal homepage such as result (a) nor an informative page from a well-regarded source, such as result (b), can claim the unique Vital status in relation to a disease query.
[How to build a fence]	http://www.theworkshop.net/Tips/htm/fence_howtobuild.htm	Good, but not unique. Many fence models out there, many opinions on how best to construct each.
[music]	http://www.music.com/ http://www.music.org	Please refer to the discussion below on URLs that match the query verbatim.
[quality of life],	http://www.utoronto.ca/qol/	A concerted effort to research the quality of life cannot speak on this query with unique authority.
[wrongful dismissal from employment],	http://www.wrongful-dismissal.com/	Although this site is solely concerned with the wrongful dismissal cases, it isn't a uniquely authoritative resource for the body of law on wrongful dismissal.
[london student apartments]	http://www.londonnet.co.uk/ln/guide/accomm/budget_student.html	A good list, but again, unless, counterfactually, all student apartments in London are monopolized by one agency, no unique resource exists for this query.

Please note: Certain queries for familiar named entities will jump at you as such. You would be able to tell that those queries have Vital results without pressing a key or pointing a mouse. Others may wrongly appear generic. For instance, [interact australia] may appear to be an awkward query placed by someone in need of online or offline companionship on the continent. However, doing web research for the query quickly reveals that there has long been a unique organization by the name of "Interact Australia". Similarly, [men's health online] could be broadly generic, but given that there is a magazine called exactly Men's Health, the likelier scenario for the query is that it's targeting the online version of that magazine. Even more likely does the query [economist] pertain to the magazine, making Economist.com Vital to the query. Why? Existence of the magazine, and now the online resource for it, gives this single word query a very strong, dominant interpretation. What about the generic interpretation? It is weak. Taken the generic sense, it's unclear what [economist] would look for – information on economists? On the latest Nobel Prize award in economics? Schools, professional organizations with directories of economists? Economist jokes? When the query has a vague generic meaning and a clear named entity

meaning, treat the generic interpretation as a minor one at most. By this token, appropriate results to [amnesty], ["Amnesty International"] and [amnesty international] queries should be the same, as the queries are essentially no different from one another – the generic meaning of [amnesty] is weak.[8]

To [legal information institute], http://www.law.cornell.edu/ is Vital because it is the homepage for a widely (and internationally) cited resource, LII at Cornell Law School. It is THE resource that most people who could have placed the query (legal scholars and practitioners, students of law, legislators, anyone interested in legal research) would associate with the query. At the same time, there is an international network of legal information institutes, to which the LII at Cornell belongs. Existence of the network does not render the query generic: the network and its non-US branches are less known. It's inconceivable that the representative user behind [legal information query] would be aware of the network or of a regional institution, such as the Australasian Legal Information Institute, without knowing of the Cornell Law School LII. Hence, had the (rational) user wanted the British or Australasian resource, the query would have reflected that preference. The network homepage and regional institutes merit high ratings but are not Vital.

What if an ambiguous query has two strong interpretations, so that each can be roughly assigned the probability of 50 percent, *and* each interpretation "possesses" a unique homepage? Do we have a Vital result for each? The answer to this question is no. Reflecting the ambiguity inherent in the query, we demote what otherwise would have been a Vital result to the next rating down the merit scale. As a result, both unique, ideal results – one per interpretation – should be rated Useful.[9]

Similarly, if one interpretation of the query happens to have a uniquely matching homepage, but the interpretation does not stand out as the most salient, predominantly likely one, then the homepage which would have been Vital in the absence of other, stronger query interpretations, should be appropriately demoted on the merit scale. In essence, you as a rater will make two judgments before arriving at the final rating for results to ambiguous queries: first, you will determine the rating applicable per interpretation. Second, you will map this rating onto the merit scale considering the presence and relative likelihoods of other query interpretations. Then, what is Vital to the dominant interpretation becomes the final Vital score, but what would have been Vital to a non-dominant interpretation should be mapped down on the merit scale,

To sum up, a Vital result is one that uniquely matches the most dominant query interpretation.

---

[8] In print, an all-lowercase string is less likely to identify a named entity than one in which initial letters are capitalized. However, it is well known among web users that Google does not distinguish upper- and lowercase in queries; hence a query without caps might just reflect efficiency on the part of an experienced user.

[9] Please see the [ADA] example below for the application of this rule.

What if the URL matches the query verbatim? Doesn't it make it Vital?

We evaluate the page and not the URL, although the URL information can be taken into consideration among other attributes of the page.

It's important to realize that sites on {genericsubject}.com or .net or .org domains may aspire to the status of an all-encompassing resource on {generic subject} but are, at best, Useful. 'Art' is not a named entity; no site, no matter how comprehensive, can claim the unique status of authority on everything related to art. Owning the www.wine.com domain does not amount to making the word wine a trademark, or to owning the body of knowledge about wine, or to exclusive rights to transactions in wine. With 652 cheeses in the database at www.cheese.com, the page may be a terrific resource in response to the query [cheese] – but this fact still does not warrant promoting the result to the very special status of Vital. Distinguishing between queries that are generic in the most salient interpretation, on the one hand, and non-generic, named entity queries, on the other hand, is an important starting point in researching coverage for a query.

USEFUL

Useful is our next category, below Vital on the merit scale. For generic queries, which do not have Vital matches, Useful results are very good results that deserve high positions in an idealized search result rankings. They are "as good as it gets", at least along one important dimension. Their attributes are constructive comprehensiveness, quality, precision in "answering" the query just right – neither too broadly nor too narrowly, authoritativeness, timeliness. It is not necessary, though, for a Useful result to possess all of the above attributes. In fact, it may not be possible: for instance, the most comprehensive book on a queried individual deserves to be called Useful because of the depth of coverage it provides. However, the book certainly cannot incorporate the news of the day on that individual on an ongoing basis. Conversely, a news site may be Useful if reliable and timely without offering the benefit of great depth.

Useful results ought to be highly satisfying for the user: if the query is informational, they should be very informative; if the query is transactional, they should allow the user to complete the transaction.[10] Table 3 contains a few examples.

[10] If you happen to have expertise in the knowledge area covering the query, and a result strikes you as meriting high position in the idealized search result ranking but you cannot exactly pinpoint what it is that makes it Useful, go with your intuition. As an expert on the area, you are in an enviable position to evaluate results from a point of view similar to that of actual users.

Table 3.

Query	Result URL	Description of the result	Appropriate Rating
[FREEDOM OF INFORMATION ACT]	http://archive.aclu.org/library/foia.html	Very helpful and informative guide from an established independent source. While this result may be more helpful to most people who need to use the Act, it's not Vital: ACLU is not a legislative authority.	Useful
[West Nile Virus]	http://www.cdc.gov/ncidod/dvbid/westnile/	Informative, authoritative page.	Useful
[West Nile Virus]	http://westnilevirus.nbii.gov/	Less helpful to the majority of people, but still authoritative and informative.	Relevant (Useful also acceptable)
["apple pie" recipes]	http://www.recipesource.com/desserts/pies/index3.html	Try some! The recipes are professionally indexed for convenient use. The resource has more recipes than most other apple pie resources you can find.	Useful
["apple pie" recipes]	http://eat.epicurious.com/	Probably the most distinguished online recipe collection. Before rating, you must search the site using the search box or advanced search to ascertain that there are enough recipes of interest (Epicurean focuses on gourmet recipes while apple pies are a more mundane fare)	Useful if out of 16,000 recipes enough are on apple pies; Relevant if only a few are Also, compare to web coverage for the query – there are probably enough apple pie devotees to demote anything that's not stellar to Relevant
[GRE]	http://www.800score.com/gre-index.htm	Helpful hints for test-**t**akers (Vital is the homepage on the site of the test-**m**aker, the ETS http://www.gre.org/splash.html)	

For queries that can have unique homepages, Useful results, too, merit high positions, though in an idealized ranking they should come after Vital pages. Useful results to homepage queries may be:

- Certain pages on the correct *site* but not the unique target of the query. E.g., the "download" page on the site of company XYZ that offers popular applications is Useful to {company XYZ} search, while the homepage of XYZ is Vital. Similarly, the store locator page on the site of a large store network can be Useful. Or consider the hours and location page if the queried entity is one that people likely visit in person.

- Results that are Useful to the non-navigational aspect of the query. For example, [simon and garfunkel] query may well be placed not just in expectation of the homepage of Simon and Garfunkel, but in expectation of browsing through good resources for tablatures, fan sites, books, etc.

- A homepage that for a query interpretation when the query has a 50:50 split between two interpretation, in other words, the homepage that *would have been* Vital in the absence of a "competing", equally likely interpretation. Demoting the result to Useful from Vital in such a case reflects not the absence of the Vital attributes of the result, but the uncertainty regarding multiple intentions behind the *query*. For instance, take multiply-ambiguous acronyms such as the following:[11]

    - ADA can stand for an important body of law, the American with Disabilities Act, of for one of at least a handful well-known associations, i.e.American Dental Association, American Diabetes Association, the American Dietetic Association. All these homepages are Useful to [ADA].[12]

    - Unless you are an editor and a new parent at the same time, you may not know that AAP stands for the Association of American Publishers and for the American Academy of Pediatrics. And even practicing pediatricians may not be acutely aware, on a day-to-day basis, of the existence of the American Academy of Periodontology…

- A recent news article about the object of the query.

---

[11] Although acronyms can often mean several things, users often place acronym queries without "supporting documentation", be it because they operate under the assumption that the organization or concept they know is the only one under the sun, or be it because they aren't certain about the unabbreviated name of what they are looking for.

[12] Note: if an acronym has a universally recognized meaning (think CIA and FBI for the Unites States), results that match an arcane de-acronymization, if such exists, should be rated very low. This is because there is no uncertainty. A reasonable person researching Cardiovascularly Insupportable Attendee ™ or Familiarly Belligerent Intelligibility, Inc. better spell those terms out, even if in a very limited circle of the initiated, these objects of search are commonly referred to by acronym.

- A homepage that should be demoted from Vital based on geographical considerations may be Useful (though often it is no more than Relevant, at best):

  - A regional homepage of significant importance when the region is not specified by the query. E.g., to [amnesty international], http://www.amnesty.org/, the global homepage, is Vital. www.amnestyusa.org, www.amnesty.org.uk are Useful.

  - A global homepage when the region is specified, or when the queried entity's activities are mostly contained in one region. Such global homepage can be also Relevant. E.g., {company primarily operating in New Zealand} is best matched by its New Zealand homepage. However, if there is also a global homepage with links to offices in a few other countries, such homepage can be Useful. To [ikea Canada], www.ikea.ca is Vital, and global www.ikea.com, Useful.

Some queries presuppose directories, i.e.collection of links, as their best results. Often but not always, queries with a plural noun ([recipes], [maps], [US embassies in Europe]) "ask" for lists. For example, to the query [newspapers in Scotland], an annotated listing of newspapers published in Scotland, http://www.wrx.zen.co.uk/scotland.htm, may have higher utility than the homepage of any individual newspaper. Of course, to have true utility the collection of links must be working – for you as a rater this means that you will have to check several links to confirm that they function.[13]

Other queries are best matched by a page with a searching functionality. In essence, finder pages offer a convenient way to search a large database. For a sample, for the [weather in {location}], a reputable weather resource that, if searched, delivers the forecast for the location, might be Useful. So would be the page on the site of the museum of {location} history that details the weather trends at {location} for several decades. Once again, results can be Useful along different dimensions… The same is true of other categories on the utility scale

To sum up: Useful are good, yet not uniquely authoritative, resources. For most queries, a result that's "as good as it gets" is Useful.

RELEVANT

One notch down the scale, Relevant results will have fewer valuable attributes than Useful results for any given query. Because our discrete categories attempt to capture what in reality is a continuum AND is open to subjective differences in opinion, you will at times

[13] **Very important!** While the cache and the live page may appear the same, their functionalities often differ: links, images, animation may be disabled on the caches. **Unless the cache materially differs from the live page, checking the links must be done from the LIVE page.**

find yourself vacillating between two possible ratings even after thoroughly examining the result in light of the query coverage. In those cases, please go with the lower rating.[14]

Relevant results may be: a single lamb chop recipe to the query [lamb chops], an amateurish personal page of a fan of a queried music band, one model of ski boots for sale where comprehensive resources for [ski boots] exist, a brief newspaper obituary on a queried politician.[15]

A listing of ski boots to a query {ski boot by precise model number}may be Relevant if it contains a link to a page with the "correct" pair of boots. In general, note that , a comprehensive resource is only then Useful when breadth is requested; in case of a query for a specific product model, a long listing covering various models conveys lack of focus – a mismatch between query and result – and deserves a relatively lower rating.

We saw already that queries that possess Vital homepages can fetch Useful results. They can also bring up Relevant results, for instance, less important pages that are on the correct *site*. [16]

A Relevant result may cover one important facet of the query only, whereas a Useful result is expected to cover the query more broadly, more thoroughly. Oftentimes, a result that organizes a vast body of information - as the table of content does in a book - is Useful, whereas individual information pieces – individual chapters, using the analogy with the organizational structure of a book -- should be assessed as Relevant or lower. (Then again, a very important "chapter" might still be Useful. Please realize that our categories are broad, and a result that's "somewhat" worse than another can very well fall into the same category with the better result.

As an example, for the query [FREEDOM OF INFORMATION ACT] http://archive.aclu.org/library/foia.html is a Useful result. http://archive.aclu.org/library/foia.html#request , on the same site, is also helpful but not as broad. It's Relevant.

Finally, Relevant is reserved for a homepage that *would have been Vital* had a more dominant query interpretation not overshadowed the minor interpretation that's matched by the homepage.

For example, [stairway to heaven] http://www.thecrowsloft.com/crowtv/, the homepage for the show titled "The Stairway to Heaven," is Relevant because the TV show interpretation for this query is subordinate to the dominant ("the" song by Led Zeppelin) interpretation of the query.

---

[14] As an exercise, research the query coverage for queries in Table 3 above and decide for results that are mentioned both as Useful and as Relevant, which rating category is most appropriate – in light of the query coverage, of course.

[15] Unless the politician just died, in which case the obituary will be Useful for a short period of time.

NOT RELEVANT

Further down on the utility continuum, Not Relevant results are generally not helpful to users but are still connected with the query topic: you can see a relationship, albeit an attenuated one, between the query and the result. Thus, on-topic results that are too marginal in scope, outdated, too narrowly regional, too specific, too broad, etc. are Not Relevant.

Take the query [yellow pages]. Of a hundred English-speaking users who pose this query, how many do we expect to be based in New Zealand, statistically speaking? Very few. Hence, the New Zealand Yellow Pages should not be in the top ten results. They are Not Relevant.

Consider now a broad informational query: [information on law school programs]. The query is clearly placed with an expectation of a broad resource. Technically, information on the sports law program at Tulane Law School in New Orleans, http://www.law.tulane.edu/prog/index.cfm?d=specialty&main=specsport.htm, fits the query – it does describe **a** law school program. However, it's too narrow on several levels – it's an isolated page from one law school covering one specific, esoteric program. It's Not Relevant. To provide some concrete meaning to the Not Relevant category as applicable to informational queries, think of someone who would want to do an extremely thorough, exhaustive research on the topic of the query, starting from the core resources and methodically expanding outward towards more marginal resources. Somewhere there, close to the perimeter of the topic, is the Tulane sports law program. Also on the edge, even further away from the central resources on the topic, is the Law and Economics Association of New Zealand, http://www.leanz.org.nz/: another Not Relevant. To the majority of users such results offer *zero utility*. Likewise, for the query [cell phone plans pricing Germany], only current results are helpful from the majority's standpoint. An outdated result is impractical and is not helpful to most users – but imagine an archivist tasked with the project to tabulate "Global trends in cell phone pricing since time immemorial", and the result's value is resurrected.

Or consider the query [foods containing sodium]. Results on sodium content of any single food would be of significantly lesser interest to the likely user "behind" the query than would be an authoritative dietetic resource covering a comprehensive list of foods.[17] Still, a "how much salt there is in peanut butter" page is marginally related to the topic of the query, whereas a "how much fat there is in peanut butter" page isn't – it's Off Topic.

A bare-bones transactional page to {product by type, reviewed} is Not Relevant because it covers the query only marginally: it lacks the requested reviews.

It is that minor, marginal interest, the tangential relationship of a result to the topic of the query that the Not Relevant category captures.

---

[17] The user here is probably a hypertensive patient, as subject research will quickly tell you.

When dealing with queries consisting of several words, it is sometimes helpful to distinguish between the exact topic of the query and the general theme. For instance,the query [maps world of warcraft] has an obvious specific topic and a general theme: the game "World of Warcraft". Only results with information on the specific topic can be rated as Relevant+ (or, indeed, have Vital matches if the developer of the game provides the maps). Pages about World of Warcraft but not on maps specifically only fit the general theme; they do not fit the exact topic of the query and hence are Not Relevant.

Another very important subcategory of Not Relevant results are pages that link to good results *without being good results themselves*. One example is a useless subpage of a Vital *site.* The errata page on the site of the New York Times, http://www.nytimes.com/corrections.html, is Not Relevant to the query [New York Times]. Reason: the resulting page is much too narrow/specific, and only a link to the page that matches the query is provided. Yes, there is easy navigation from this URL to the most likely query target, the newspaper's homepage, but outside that link there is no utility here for the overwhelming majority of users.[18] By contrast, the editorial page and the subscription page of the paper have information value in themselves to most users, besides offering a link to the top level homepage; hence, they may be Relevant or Useful.

Note: If navigation to helpful content is difficult from the "wrong" page on a good site, please feel free to demote such a result further down to Off Topic category discussed below.

The same goes for a result page on one site that links to a Vital or Useful result on another site without providing any utility in and of itself, other than the link. For the query [Library of Congress], the listing http://www.indiana.edu/~librcsd/internet/libweb-mirror/usa-org.html has little utility in and of itself. It contains a link to the homepage of the LOC among many other entries, but the user who needs to access the Vital homepage should be able to do so directly, without changing planes in Indiana. As a link away from the target, this result is Not Relevant. Sure for an archivist who'd want to count how many libraries worldwide link to LOC, the stop in Indiana may be justified. For the rest of us, it's not.

OFF TOPIC

Further down the merit scale is the Off Topic category. Off Topic results would not even interest our hypothetical archivist and would not even fit a query interpretation that's minimally plausible. They are not even tangentially related to the query and are of zero utility.

For certain queries, there is a wide gap between Not Relevant and Off Topic results. For example, to [coffee grounds] query, a resource on penguin species is totally random and most definitely Off Topic. However, because Google is very literal in matching query and

[18] Barring an unlikely scenario that the Times should commit a typo that becomes the talk of the day.

result text, instances of such out-of-the-blue results are few and far between. More widespread are instances where there is a matching word on the result page, but the page, due to contextual factors, is still Off Topic. For instance, that http://www.penguin.co.uk/ is Off Topic to [penguin species] query is crystal clear.[19] And even more frequent are Off Topics that have a very strained relationship to the query but do not merit the Not Relevant label. It's on those results that reasonable raters can disagree, and it's those results for which there is a continuum between Off Topic and Not Relevant.

A frequent instance of Off Topic results is a page with query words occurring on different frames, or in different places in the text, unrelated to each other.

Another instance is a result (returned to a query composed of several words) that matches only the less salient, less definitive word(s) in the query. Or, one that matches one keyword but strips it of the context provided by the other word(s) and thus crucially changes the meaning. What renders a result Off Topic is lack of attention to a restricting modifier in the query. Table 4 shows several examples where this applies.

Table 4.

Query	**Resources that *may* be Not Relevant, although Off Topic rating too is appropriate**	**Resources that are clearly Off Topic**
[aromatherapy classes in Bromsgrove]	Aromatherapy resources in general; online aromatherapy classes	Bromsgrove museum In Bromsgrove, UK
[English to Latin Translation]	Resources on Latin other than translation from English to Latin, in particular Latin to English translation	Translation resources for English to modern languages
[berkeley empiricism]	Results on philosophy without overt mentions of Berkeley's contribution	Any result related to UC Berkeley
[world map {telling omission}] By *not* specifying *which* imaginary world, the query defaults to the accepted, geography of the Earth, meaning of the queried word combination.	Good geography resources that allow the user to retrieve maps of large regions, without giving an option to see the map of the world	Maps pertaining to imaginary worlds of online games, such as http://heroes.net.ru/map-erathiaen.shtml

[19] Unless, of course, the publishing house site features a book about the penguin species.

Table 4 (cont'd).

Query	**Resources that *may* be Not Relevant, although Off Topic rating too is appropriate**	**Resources that are clearly Off Topic**
[pampers {telling omission}]	Resources on the financial position of Procter and Gamble with a passing mention of the Pampers brand of diapers	- "From Pampers to Depends" book for sale: given the popularity of Pampers diapers, it's not conceivable that anyone would ever search for the book by querying [pampers] without further a do.[20] - A result page with the word 'pampers' in the generic sense (as in "person A pampers person B")

As with other decisions on the utility scale, the Off Topic versus Not Relevant decision should be ultimately resolved in favor of the lower rating – Off Topic. Why? If you are unsure whether the result deserves Not Relevant, it probably does not.

IMPORTANT: some queries are defined so tightly that no broadening or narrowing the topic is possible. To those queries, results are either quite good or Off Topic: none or very few Not Relevant results are conceivable. That is OK – we cannot say it often enough, go with your best judgment and do not worry too much about rationalizing every single rating decision. What we hope to instill via these Guidelines is a general understanding of the rating methodology. Once you internalize the criteria for placing results on the merit scale based on attributes of the result, the query, and the query's web coverage, you will be able to apply the scale to types of cases not covered in these Guidelines either because they were not foreseen or because they were intentionally omitted from the discussion for the sake of brevity.

OFFENSIVE

As with other names of the categories, this one has a dictionary meaning that does not necessarily mesh with the category's technical meaning.

[20] Had the book being widely acclaimed, one could change the opinion on the proper rating of the book result. But no, check http://www.nytimes.com/pages/books/bestseller/, it's not a bestseller.

Offensive results are at the very bottom of the merit scale. They are not on the utility continuum; in many ways, determining whether a result is Offensive is orthogonal to utility considerations.

Offensive results very often are Offensive independent of the query – that is, they do not have merits for *any* query. If a result attempts to wreak havoc on your computer (load a worm, create a loop that necessitates closing all browser windows, etc.), there is no query (save for a query that uniquely targets the result page via a clear, one-to-one correspondence) to which this result is any good. If a result displays evidence of cheating techniques, for instance if it's a page created for search engine robots rather than for human visitors, such result does not deserve to be brought up anywhere in the ranking to most queries, except, once again, to queries specifically targeting the page.

Other Offensive results are offensive in a less absolute way – they are offensive to some queries and not to others. For example, *uninvited* porn results are definitely Offensive. Yet, some queries "invite", and others "tolerate", porn results. While we remove explicitly pornographic queries from our query sets, we retain queries with various nuances in meaning, some of them more "adult" than others. Queries such as [high boots] and [nylons alexandra] will serve for an example. Query coverage to many a software download query happens to "reside in a pornographic neighborhood"; demoting pornographic results to such queries would effectively limit the set of possible ratings for results to those queries to one single rating: Offensive. Since doing so will not send any meaningful feedback to the engine, we ask that you not label results Offensive based on *your subjective perception*, but put yourself in the shoes of a representative user on a per query basis.[21]

A frequent application of the Offensive label is to results that fall under the category of web spam (deceitful web design). To give you a flavor of what spam results can look like, we offer several examples in Table 5. These examples do not purport to cover the topic of spam in its entirety. A separate document, Spam Guide, focuses exclusively on spam identification tools and is required reading after you become comfortable with these General Guidelines.

[21] If you object to pornographic web environment, we will accommodate your preferences and not require that you rate objectionable content.

Table 5.

Query	Result URL	Explanation
[Charleston, SC Chamber of Commerce]	http://www.jicccharleston.com/charleston-sc-chamber-of-commerce.shtml	This site is set up solely to get money from clicks to sites that it links to: it gets paid for every click on the link.
[laetitia casta]	http://www.laetitia-casta.com/	Checking the Properties of this page, we see the true URL address: http://ww2.sextoysex.com/sex/start/sex.html?a=227 This sex toy shop attempts to lure visitors via a pretense of high relevance to the Laetitia Casta query.
[Learn How To Knit]	http://www.searchresults.ws/how-to-knit.htm	This result is an example of "secondary search result" type of spam, discussed in detail in Spam Guide for Raters.
[photographers in Hawaii]	http://www.anthonycalleja.com/	Anthony Calleja is a photographer in Hawaii. His page would not have merited a high rating to this query since the query most likely asks for lists, but it would not have been a totally useless result either. The photography business in Hawaii is a highly competitive one. In an attempt to get ahead of his competition by promoting the site to higher ranking positions in various search results, Mr. Calleja's evidently retained services of a "spammer" webmaster who stuffed the page with thousands of popular searches related to photography, weddings, modeling, Hawaii, and other, most general terms. These keywords are not visible to the human visitor to Mr. Calleja's page (but you can see them by clicking ctrl-A in Internet Explorer). Note that this page should NOT be rated Offensive to the query [Anthony calleja].

IMPORTANT: Observe the query-matching words in the URL structure of the first three examples. Spam tactics such as these ones is another reason not to take URL addresses at face value, but to evaluate actual pages.

Concluding Remarks

As you see, the rating task consists of

- Understanding the meaning of the query and its type – is it navigational, informational, transactional, or a mixture of two or three?

- If you come to the realization that the query could have been posted by different users with different intentions, crudely assigning possibilities for each interpretation and/or intent

- Researching the query coverage on the web using search engines other than Google, directories, specialized databases, and other sites, or offline resources[22]

- Examining each result for attributes that would call for assigning an applicable special category rather than a merit-based assessment, and, in the absence of those attributes,

- Determining the merit rating in light of the query coverage and considering various utility dimensions, as well as taking into account evidence of deceitful web design where appropriate.

[22] Research for the query should be done before you open any results that are up for evaluation