General Guidelines

Version 2.1 (April 6, 2007)

Part 1: Rating Guidelines

Welcome to the Quality Rating program! If you have previously read Version 1 of these guidelines, please pay special attention to the text highlighted in yellow. If you are new to the program, please read the entire document very carefully.

1. The Role of the Quality Rater

As a Quality Rater, you will evaluate '**query-page**' Tasks. For each 'query-page' Task, you will:

- Research and understand the **query**.
- Evaluate the **page** based on its relevance to the query and its utility to the user.
- Assign a rating from the Rating Scale.

Query refers to the word or words that a user types in the search box of a search engine.

The **URL** is the web address of the page you will evaluate, such as http://www.microsoft.com/.

The **Page** or **Landing Page** is the page you will evaluate. It is the page you see after you click on the URL.

Task Language and Task Location. You will be given a **Task language** and a **Task location** for each query-page Task. You must evaluate each Task in the context of its language and location.

In this document, each query will be shown in square brackets, followed by the Task language and Task location. Examples:

[Elvis Presley], English (US)
[coca cola], Spanish (MX)

Please keep in mind that the language of the query may not match the Task language. For example, you may be working on a German (DE) Task and see a query in English.

2. Researching and Understanding the Query

You should understand each query before you evaluate it. If the meaning of the query is unclear, you will need to do web research to learn about it. You can do this by entering the query in the search box of one or more search engines and looking at the results returned by them. However, your rating should not be affected by the ranking of results you see displayed by the search engines.

You will also need to understand the possible interpretations of the query and try to imagine a user who would type the query. Think about what the user might be trying to accomplish.

Here are some things to consider:

Task Language and Task Location

All queries have a Task language and Task location. You must use the Task language and Task location to understand the context of the query. Users in different parts of the world have different expectations for the same query terms. Imagine the user typing in the following query and what they would be looking for.

Query	Task Language	Task Location	Query as Typed by the User	Possible User Expectation
[George Bush]	English	US	A user in the United States types the query [George Bush].	George Bush's official government web page
[□□□□]]	Chinese Traditional	Taiwan	A user in Taiwan types the query [George Bush] using Traditional Chinese characters.	Information about George Bush displayed in Traditional Chinese
[George Bush]	Chinese Traditional	Taiwan	A user in Taiwan types the query [George Bush] in English.	George Bush's official government web page displayed in English

The query may have different meanings, depending on either the Task Language or Task Location.

Query	Dominant Interpretation in the Task Location
[football], English (US)	The dominant interpretation is American football played with a brown oval ball.
[football], English (UK)	The dominant interpretation is the game Americans call soccer and which is played with a round ball.

Multiple Interpretations

Does the query have more than one interpretation? Try to imagine the possibilities. Is one interpretation the most likely or dominant interpretation?

Query: [windows], English (US)	
Dominant Interpretation: A universally known computer operating system	**Possible Interpretation**: A piece of glass that can be looked through

Query: [java], English (US)		
Dominant Interpretation: A programming language	**Possible Interpretation**: An island in Indonesia	**Possible Interpretation**: Coffee

Query: [mercury], English (US)		
Possible Interpretation: The planet	**Possible Interpretation**: The chemical element (Hg)	**Possible Interpretation**: The car

Broad or Specific

Is the query broad or specific? Broad queries are best matched by broad pages; specific queries are best matched by specific or narrow pages.

	Broad	Specific
Query	**[digital cameras]**	**[canon SD550]**
Possible intent	*Looking to purchase a digital camera.*	*Looking to purchase this specific camera*
Well-matched result	http://www.bestbuy.com/site//olspage.jsp?id=cat04001&type=category *Broad result for a broad query. (good)*	http://www.amazon.com/Canon-Powershot-SD550-Digital-Optical/dp/B000AYKUUQ *Narrow result for a narrow query. (good)*
Poorly-matched result	http://www.bestbuy.com/site/olspage.jsp?skuId=9999502000500048&type=product&productCategoryId=pcmcat99300050011&id=pcmprd50400050004 *Narrow result for a broad query. (not so good)*	http://www.amazon.com/s/ref=nb_ss_p/103-3349756-5881468?url=search-alias%3Dphoto&field-keywords=digital+cameras&Go.x=0&Go.y=0&Go=Go *Broad result for a narrow query. (not so good)*

Amount of Information Available

If there is a lot of information available on the Web for the query, then a page with just a link or a short article is not a good search result. If there is very little information available, you may consider the page to be a good result.

Timeliness

Can a query be interpreted differently at different points in time? In 1994, the user who typed [President Bush], English (US) was looking for information on President George H.W. Bush. In 2006, his son George W. Bush is the more likely interpretation. You should always rate according to the current interpretation or in the appropriate context if the query has explicit date information.

3. Query Types

Most queries can be classified in one or more of the following three categories: Navigational, Informational, or Transactional.

Navigational

A **navigational** query is intended to locate a specific web page. The user has a single web site in mind, often the official homepage or subpage of an official site. For example:

Navigational		
Query	**URL of the Landing Page**	**Description of the Landing Page**
[ibm], English (US)	http://www.ibm.com/	Official homepage of the IBM Corporation
[yahoo mail], English (US)	http://mail.yahoo.com/	Official Yahoo! Mail web page
[ebay], English (US)	http://www.ebay.com/	Official homepage of eBay
[ebay], Italian (Italy)	http://www.ebay.it/	Official homepage of eBay Italy
[Harvard libraries], English (US)	http://lib.harvard.edu/	Subpage of an official site

Informational

An **informational** query seeks information on a topic. The user is looking for information on the query topic (broad or specific). The goal is to learn something by reading or viewing content on the web, such as text, images, video, etc.

Informational		
Query	**URL of the Landing Page**	**Description of the Landing Page**
[tsunami], English (US)	http://en.wikipedia.org/wiki/Tsunami	Wikipedia site with comprehensive information
[Switzerland], English (US)	https://www.cia.gov/cia/publications/factbook/geos/sz.html	Informative CIA web page on Switzerland

Transactional

A **transactional** query seeks to complete a transaction on the Web – for money or free – of a product or service. The user is mainly looking for a resource (NOT information) available via web pages. The goal is to download, to buy, to obtain, to be entertained by, or to interact with a resource that is available on the result page or made available through the result page.

Transactional		
Query	**URL of the Landing Page**	**Description of the Landing Page**
[Beatles poster], English (US)	http://www.allposters.com/-sp/-Posters_i317216_.htm	Page on which to purchase poster
[download adobe reader], English (US)	http://www.adobe.com/products/acrobat/readstep2.html	Official free download page on Adobe website

Source: http://vizualbod.com/f/spam-guidelines.htm - 12/03/2008

Many queries fit into more than one category. For example: [united nations], English (US). The user may expect to be taken to the homepage of the United Nations: http://www.un.org/ (navigational intent), or the user may be looking for recent news regarding a United Nations resolution (informational intent).

Queries that fit into more than one category			
Query	**Navigational Intent**	**Informational Intent**	**Transactional Intent**
["ipod nano"], English (US)	Looking for the official product page	Looking for reviews or product information	Looking to purchase the product
[britney spears], English (US)	Looking for the official website of this celebrity	Looking for pictures, latest news, forums, etc.	Looking to purchase a poster or download music or video
[united nations], English (US)	Looking for the official website	Looking for recent news regarding a United Nations resolution	None
[cloth diapers], English (US)	None	Looking for information about using cloth diapers	Looking to purchase the product

4. Rating Scale

After researching the query, you will evaluate the page that loads after clicking the link provided. Most pages will be assigned a rating from the Rating Scale: **Vital**, **Useful**, **Relevant**, **Not Relevant**, or **Off-Topic**.

Vital

The **Vital** rating is used in the following special situations:

Query: The query has a dominant interpretation. The dominant interpretation is navigational.
Page: The page to evaluate is the official web page of the query.

Here are some examples:

Vital Results for English (US) Queries			
Query	**Vital**	**Vital Page URL**	**Description**
[Singapore airport]	Yes	http://www.changi.airport.com.sg/	Official homepage
[toyota camry]	Yes	http://www.toyota.com/camry/index.html	Official product page on the correct site
[apple]	Yes	http://www.apple.com/	The dominant interpretation of this query is Apple Computer, Inc., and this is the official homepage.
[barnes and noble]	Yes	http://www.barnesandnoble.com/ http://www.bn.com http://www.books.com	Multiple **Vital** URLs with the same landing page (different domains with the same owner)
[Fleet Bank]	Yes	http://www.fleetbank.com/index.cfm http://www.bankofamerica.com/index.cfm	Multiple **Vital** URLs with the same landing page. Fleet Bank was acquired by Bank of America in 2004.
[yahoo]	Yes	http://www.yahoo.com http://www.yahoo.com/?200204a	Multiple **Vital** URLs with the same landing page (same domain)
[download firefox]	Yes	http://www.mozilla.com/firefox/	The download page on the official site
[Hillary Clinton]	Yes	http://clinton.senate.gov/	Hillary Rodham Clinton's official U.S. Senate page
[knitting]	No	none	This is an informational query
[ipod reviews]	No	none	This is an informational query

Source: http://vizualbod.com/f/spam-guidelines.htm - 12/03/2008

[ADA]	No	none	There is no dominant interpretation. The following are all strong interpretations: American Dental Association American Diabetes Association American with Disabilities Act All of these interpretations have official homepages, but none is **Vital**.
American Dental Association	Yes	http://www.ada.org/	Official homepage of this organization

Most queries do not have **Vital** results because they are not navigational, and/or do not have official web pages or official product pages associated with them.

All queries have a Task language and a Task location. For a landing page to be **Vital**, it must be appropriate for the Task language and Task location.

Vital pages may not be the best possible result for the query. In fact, it is possible for a **Vital** page to be not the most helpful at all, for example, in the case of a celebrity homepage that does not have what the user wants to know.

The **Vital** rating is not based on the appearance or content of the URL. Often, the URL of the official homepage will contain the query terms. For example, the **Vital** result for [ibm] is http://www.ibm.com.

However, sometimes a URL contains the exact query terms, but the landing page is not **Vital.** For example, www.diabetes.com cannot be **Vital** for the query [diabetes] because the query is not navigational and there is no official web page for the query. No person or entity can claim ownership of the query [diabetes]. If no one can "own" the query, there can be no **Vital** result.

You will often see URLs that contain celebrity names, but when they are not maintained by the celebrity, they are not official homepages and are not **Vital**. For example, http://www.jenniferlopez.com/ is **Vital**, but http://www.jenniferlopez.net/ is not. Web research is needed to determine whether a landing page is **Vital**.

Some large international corporations have country, as well as regional or global homepages. In general, the country specific homepage is the **Vital** result for that type of query. If no country specific homepage exists, a regional or global homepage may be **Vital**.

Sometimes the landing page asks you to choose a language, country, postal code, zip code, etc. These pages should receive a **Vital** rating, if the pages behind them would receive a **Vital** rating. Similarly, splash and flash pages should also receive a rating of **Vital**.

It is not uncommon today for individuals to maintain various types of personal pages on the Web. Homepages, social networking pages, and blogs have become increasingly popular. Some individuals have more than one blog and/or more than one homepage on a social networking site (e.g. myspace, facebook, friendster, mixi). When these pages are maintained by the individual (or an authorized representative of the individual), they are all considered to be **Vital**.

Useful

A rating of **Useful** is assigned to pages that contain some or all of the following characteristics: highly satisfying, comprehensive, high in quality, and authoritative. **Useful** pages answer the query just right; they are neither too broad nor too specific. For many queries, they are "as good as it gets."

Examples of **Useful** pages include: a page that is highly informative; a timely and informative article; a page that allows the user to complete the intended transaction; an important subpage on the correct site; the homepage of the correct site when a specific product is asked for. If a query "asks" for a list, then a directory (a collection of links) can be **Useful**, e.g. [fudge recipes], [books about sharks].

If an ambiguous query has several equally strong interpretations and each possesses a unique homepage, the homepages would all be assigned a rating of **Useful.**

Source: http://vizualbod.com/f/spam-guidelines.htm - 12/03/2008

Useful Results for English (US) Queries		
Query	**Useful Pages**	**Explanation**
[Microsoft]	http://www.microsoft.com/windows/default.mspx	Subpage of the Microsoft website devoted to Windows
[Honda Accord]	http://automobiles.honda.com/	Homepage of correct site for the product
[ADA]	American Dental Association http://www.ada.org/ American Diabetes Association http://www.diabetes.org/home.jsp American with Disabilities Act http://www.usdoj.gov/crt/ada/adahom1.htm	Homepages for sites with equally strong interpretations
[meningitis symptoms]	http://www.webmd.com/hw/infection/aa34586.asp	Highly informative page on authoritative site
[broadway tickets]	http://www.ticketmaster.com/broadway	Reputable site on which to complete a transaction
[books on whales]	http://www.whalecenter.org/books.htm	List of books on whales

Relevant

A rating of **Relevant** is assigned to pages that have fewer valuable attributes than were listed for Useful pages. **Relevant** pages might be less comprehensive, come from a less authoritative source, or cover only one important aspect of the query.

Examples of **Relevant** pages include a page with a brief article on the topic of the query or a less important subpage on the correct site. If a query "asks" for a list, then a single item is **Relevant**. For example, if the query is [fudge recipes], a single fudge recipe is **Relevant**.

A rating of **Relevant** is also used for a homepage that would have been **Vital** if there had not been a more dominant interpretation for the query.

Relevant Results for English (US) Queries		
Query	**URL of the Landing Page**	**Explanation**
[laser printer]	http://www.pcmag.com/article2/0,1759,1835786,00.asp	Page with info. on the HP LaserJet 2840
[seoul, korea]	http://www.escortmap.co.kr/english/e_sall.htm	Page with map of the city of Seoul
[Tom Cruise]	http://www.amazon.com/Tom-Cruise-Movies/lm/JYPI5X31Y044	Amazon page that displays Tom Cruise movies for sale.
[dell]	http://www.dellmagazines.com/	Homepage for Dell Magazines. This might have received a **Vital** rating if not for Dell Computers, the dominant interpretation.

Not Relevant

A rating of **Not Relevant** is assigned to pages that are generally not helpful, but are still marginally connected with the query topic. **Not Relevant** pages may be outdated, too narrowly regional, too specific, too broad, etc. to receive a higher rating. They might have less information or come from a less authoritative source.

A rating of **Not Relevant** is also assigned to a page that has a link to good results on the same site, but is not a good result itself. It may be an unimportant or useless subpage on the correct site. (Another example of a **Not Relevant** page is one that has a link to good results on another site without providing any utility itself, other than the link to the "good" results on the other site.)

Please note that, in both of these cases, there is a direct link from the landing page to the good result. In contrast, landing pages that are search engine pages (or pages with search boxes on them, but which do not have a relationship to the query) should be rated **Off-Topic**. If the query has to be typed in a search box, the page has no relevance to the query, even if using the search box leads to relevant results.

Source: http://vizualbod.com/f/spam-guidelines.htm - 12/03/2008

Not Relevant Results for English (US) Queries		
Query	**URL of the Landing Page**	**Explanation**
[Dentist for children]	http://www.williamsteig.com/drdesoto.htm	A storybook for children about a mouse-dentist named Dr. De Soto
[japan]	http://www.jpo.go.jp/	Homepage of the Japan Patent Office
[Louvre]	http://amazon.imdb.com/title/tt0382625/	Page about the movie "The Da Vinci Code"
[SAT college board]	http://www.ocps.k12.fl.us/links.rhtml	The value of this page is in its link to the College Board website; the page is not a good result itself.
[BBC]	http://www.bbc.co.uk/dna/mbfansforum/F2154398	The "Dundee United" Fans Forum on the BBC website

Off-Topic

A page that has zero relevance to the query should be assigned a rating of **Off-Topic**.

Ratings are not based on the presence or absence of the query terms. A page that contains the query terms, but is conceptually off topic, should be given a rating of **Off-Topic**. For example, a page on doghouses is off topic for [hot dog]. Another example of a page that should be rated **Off-Topic** is one in which the query terms occur in different places on the page, unrelated to each other. Please note that a page may also be rated **Off-Topic** even if the query terms appear in the URL.

A rating of **Off-Topic** also applies when there is lack of attention to an important modifier or element of the query. For example: [universities in India]. An article about universities in Europe is **Off-Topic**.

If navigation to helpful content is very difficult, a rating of **Off-Topic** may be assigned. For example, if the link to good results is poorly-labeled or buried at the bottom of a long list of links, or if you need to click multiple times to get to helpful content, you may assign a rating of **Off-Topic**.

Off-Topic Results for English (US) Queries		
Query	**Off-Topic Pages**	**Explanation**
[Tom Cruise]	http://www.ussslater.org/signals/vol-3/ss-v3-n4.html	Page that mentions **Tom** Beeler and **Tom** Moore and vacation **cruise**s.
[hammerhead sharks]	http://www.sj-sharks.com/	Homepage of the San Jose Sharks hockey team.
["Mission Impossible"]	http://www.avma.org/onlnews/javma/jul00/s070100d.asp	Article with title "Arctic Mission: Not Impossible"
[hotmail login]	https://login.yahoo.com/config/login_verify2?&.src=y	Login page for Yahoo! Mail
[german cars]	http://www.subaru.com/	Homepage of Subaru
[Harvard Business School]	http://www.law.harvard.edu/	Homepage of Harvard Law School
[New York Yankees]	http://newyork.mets.mlb.com/index.jsp?c_id=nym	Homepage of the New York Mets baseball team
[earthquakes]	http://www.yahoo.com/	Search engine page that has no connection to the query. Even though you can issue the query in the search engine and get good results, the rating should be **Off-Topic**.

Adjustments to Ratings Based on Task Location and Page Location

It is very important to use the Task Language and Task Location to interpret the query. You will also need to use the Task language and Task location to evaluate the page. Sometimes the Task location doesn't match the country domain of the page. For example, the Task location is Spain, but the country domain of the page is Mexico (.mx).

In many cases, when there is a mismatch between the Task Location and the country domain of the page, you will need to **lower the rating** for the page. You must use your common sense and cultural knowledge to determine whether to lower the rating and how much to lower it. Do not hesitate to lower the rating to **Off-Topic** if there is a mismatch between the Task Location and country domain of the page that would make the result useless for a user in the Task Location. High ratings are appropriate for pages with high relevance and which are in the right language and right location.

Here are some examples:

[Ikea], English (US)			
Description	**URL of the Landing Page**	**Rating**	**Notes**
Homepage of Ikea ("Select your location" portal page)	http://www.ikea.com/	**Vital**	This page is **Vital** even though it is not the US page.
Homepage of Ikea US	http://www.ikea.com/ms/en_US/	**Vital**	This page is **Vital**.
Homepage of Ikea United Kingdom	http://www.ikea.com/ms/en_GB/	**Not Relevant**	This is not the target of the query for English (US) users, but is related to the query.
Homepage of Ikea Australia	http://www.ikea.com/ms/en_AU/	**Not Relevant**	This is not the target of the query for English (US) users, but is related to the query.

[Bridget Jones's Diary], English (US)			
Description	**URL of the Landing Page**	**Rating**	**Notes**
Amazon listing with lots of reviews	http://www.amazon.com/Bridget-Joness-Diary-Helen-Fielding/dp/014028009X	**Useful**	This is a good result and the Task location matches the query location.
Page on the UK Amazon site	http://www.amazon.co.uk/Bridget-Joness-Diary-Helen-Fielding/dp/0330375253	**Not Relevant**	For most book purchases, US users would use the US Amazon site. The rating should be lowered from **Useful** to **Not Relevant.**

[Cheesecake recipe], English (US)			
Description	**URL of the Landing Page**	**Rating**	**Notes**
Cheesecake recipe	http://www.cooksrecipes.com/dessert/cappuccino_cheesecake_recipe.html	**Relevant**	The ingredients and measurements are familiar to US residents.
Cheesecake recipe	http://www.domesticgoddess.ca/recipes.php?recipe=10053	**Relevant**	Even though this is a Canadian page, the measurements and ingredients are familiar to US residents. There is no need to lower the rating.
Cheesecake recipe	http://www.sofeminine.co.uk/w/recipe/r279/cheese-cake.html	**Not Relevant**	The measurements are in metrics and the ingredients are British. Few US residents could make this cake, so the rating is lowered from **Relevant** to **Not Relevant**.
Cheesecake recipe	http://aww.ninemsn.com.au/ARTICLE.aspx?id=46528	**Not Relevant**	The measurements are in metrics. Few US residents could make this cake, so the rating is lowered from **Relevant** to **Not Relevant**.

Source: http://vizualbod.com/f/spam-guidelines.htm - 12/03/2008

There are some queries which accept or even invite results from other locations. Whether you lower the rating or not will depend on the content of the page. Here is an example:

[queen of England], English (US)			
Description	**URL of the Landing Page**	**Rating**	**Notes**
Description of landing page: Official homepage of the British monarchy	http://www.royal.gov.uk/	**Vital**	This page is **Vital** even though it is not a US page.
Article in Wikipedia	http://en.wikipedia.org/wiki/Elizabeth_II_of_the_United_Kingdom	**Useful**	This is a very good article on a US website.

Important Notes on the Rating Scale

- A page is rated on its match to the *concept* of the query (i.e. how relevant or useful the content on this page is to the query), not on the presence or absence of the query terms on the page. For [Paris Hilton picture], a photo of Paris Hilton is **Relevant** even if the query terms are not on the page. On the other hand, a page titled "Sightseeing in Paris" with a review of the Hilton Hotel in Barcelona is **Off-Topic**.

- Please remember to rate the page and not the URL. You must visit the landing page and rate the content.

- The same landing page can have multiple URLs. Identical landing pages should receive the same rating regardless of the URL.

- Please go with your best judgment and do not worry too much about rationalizing every single rating decision. Once you have a general understanding of the guidelines, you will be able to apply the Rating Scale to types of cases not covered.

- Sometimes you may feel unsure which of two ratings to give: **Relevant** or **Useful**? **Not Relevant** or **Relevant**? When you are unsure, select the lower rating.

5. Non-Rating Categories

Pages that cannot be evaluated on the Rating Scale will be assigned one of the following non-ratings: **Didn't Load, Foreign Language, or Unratable**.

Didn't Load

A non-rating of **Didn't Load** applies to many different situations. The table below displays some examples of pages that should be rated **Didn't Load**. In some cases, the page loads with no visible content. In other cases, the page loads to some degree, but lacks content that can be evaluated. The list is not complete, but can be used as a guide. You will encounter situations that are similar, but don't fall neatly into one of the categories.

Didn't Load Scenarios	Description and Examples
404 Server Error	A generic browser 404 message "Server cannot be found" message
Page not found	The server loads but the particular page is not found. "Page not found" or "Article not found" message

Source: http://vizualbod.com/f/spam-guidelines.htm - 12/03/2008

Product not found	The URL shows a product ID number, but the page loads with a "Product not found" message.
Site Unavailable	"Site Under Construction" message "Account temporarily closed" message "Host your site with us" message Exception: • If the query is navigational, the result is the dominant interpretation, and there is good reason to believe that the site will indeed be developed, the page should *not* be rated **Didn't Load**.
Blank page	A completely blank page loads in the browser, and checking the source code of the page reveals no hidden content.
403 Access Error	"You are not authorized to view or access this page." "403 Forbidden Access" "You are trying to access material included in JSTOR, an online journal archive made available to researchers through participating libraries."
Login Required	The page requires login and you would rather not sign up for free or paid registration. In general, we encourage raters to sign up with free reputable online sources and actually view the content. Such reputable sites include, but are not limited to: The Dallas Morning News, Washington Post, Forbes.com, etc. Exceptions: • If the landing page is the login page for an email account, a financial institution account, a social networking site, a message board, etc., where it is likely that the user expects to land on such a page, the rating should *not* be **Didn't Load**. • Some news sources provide free access to their articles for only a limited period of time, after which the articles are archived and are available only to paid subscribers or on a fee-per-article basis. In these cases, the landing page will generally display the article title and a sentence or two from it. Do not rate such pages as **Didn't Load**. Base your rating on the relevance of the information provided to the query. • Similarly with scientific articles, the page will provide an abstract or a portion of the abstract of an article. Again, do not rate such pages as **Didn't Load**. A scientist searching for such articles may have a subscription to the journal hosting the article, or might be prepared to purchase the article.
Encoding Error / Garbled Text	The majority of the text on the page displays encoding errors. For example: □□□□□□~□ O□□□□□□@□□□~□□□□□□□ □□□ □□ □,□Ɵ□
XML / HTML pages	A page loads with XML and HTML format that makes the content difficult to view. < ?xml version="1.0" encoding="utf-8" ?> -<!-- generator="wordpress/2.0.4" -->
Server internal error	A page loads with system error messages: **Warning**: OCIParse: invalid connection 0 in **oci8.inc** on line **21** **Warning**: OCIExecute: invalid statement 0 in **oci8.inc** on line **28** **Error** '80020009'

Exceptions: Some pages appear to fall into one of the scenarios described above, but should not be rated **Didn't Load** because they have been designed using deceptive techniques or have offensive features, such as pornography text or links. (Please see section 6 on Labels, as well as the Webspam Guidelines.)

Source: http://vizualbod.com/f/spam-guidelines.htm - 12/03/2008

Foreign Language

A non-rating of **Foreign Language** is assigned to a page that loads fine, but is fully in a third language.
A **third language** is a language *other than* English or the Task language.

A landing page in a third language should be rated as **Foreign Language** even if you are personally fluent in the particular third language. For example, if you are working on a German (DE) Task and the page is in Spanish, you will rate as **Foreign Language** even if you are fluent in Spanish. For English (US) Raters, any language other than English is considered to be a foreign language.

The **Foreign Language** rating never applies to pages in English, no matter what the Task language.

Task Language and Task Location	Page Language	Is it Foreign Language?
Chinese Traditional (HK)	English	No
Chinese Traditional (TW)	Japanese	Yes
Chinese Traditional (TW)	Chinese Simplified	No
Chinese Simplified (CN)	Chinese Traditional	No
Danish (DK)	Norwegian	Yes
Norwegian (NO)	English	No
Spanish (ES)	Spanish	No
English (CA)	French	Yes

Please note, however, that you may – and often should – lower a rating for a page in a language other than the Task Language, if good pages in the Task Language and Task domain exist and make more sense for the location. Even though English is not considered a foreign language for a result, that doesn't mean that pages in English are as "good" as pages in the Task Language and from the Task domain. Often they are not as good, and ratings for such pages should be lowered.

Above all, apply common sense when you determine what the *main* language of the page is and whether a result can be used by someone who does not speak the third language.

Exceptions:

- The page contains an image that can be evaluated in spite of the language on the page. The page should be assigned a rating from the Rating Scale.
- The page has a link for download of software, and it is understandable in spite of the language on the page. The page should be assigned a rating from the Rating Scale.
- The page has a mix of languages, but there is enough text in English or in the Task language to fully comprehend the content. The page should be assigned a rating from the Rating Scale.
- The page has a "page not found" message in a foreign language. The rating should be **Didn't Load**.
- The page is in a foreign language, but uses deceptive techniques. The rating should be **Foreign Language** with a **Spam** label. (Please see section 6 on Spam labels, as well as the Webspam Guidelines.)

Occasionally, you will encounter a page that contains text that is almost fully in a foreign language, with the exception of one link to content in English or in the Task language. The link may say "Click here for English", "Continuez en Français", etc. These pages should be rated **Foreign Language**.

Unratable

Assign a non-rating of **Unratable** when, even after researching the query, you do not feel confident evaluating the page.

Source: http://vizualbod.com/f/spam-guidelines.htm - 12/03/2008

6. Spam Labels

In the “Webspam Guidelines”, you will read about deceitful web design techniques sometimes used by webmasters to increase the ways in which users or a search engine will find their sites. Based upon your understanding of the Webspam Guidelines, you will assign one of the following three Spam labels to pages that load and can be rated:

Not Spam

If you believe that a page has not been designed using deceitful web design techniques, you should assign a **Not Spam** label.

Maybe Spam

If you find a page to be “spammy”, but you don’t feel comfortable saying that the webmaster definitely designed the page using deceitful web design techniques, you should assign a **Maybe Spam** label.

Spam

If you believe that a page has been designed using deceitful web design techniques as described in the “Webspam Guidelines”, you should assign a **Spam** label.

If you choose either **Maybe Spam** or **Spam**, please include a comment explaining why.

7. Flags

In addition to the rating (and possible Spam label) you assign to a page, you will sometimes also apply one or both of the following flags to signify specific attributes or characteristics that you observe on the page.

Pornography

If the page has pornographic links, text, images, pop-ups, and ads, it should be flagged as **Porn**. This flag should be assigned even if the query – for example, [freeones] - “invites” pornographic results. Please base your rating on the landing page.

Malicious

If the page forces you to quit the browser, if there are prompts that keep coming back, if there are attempts to download spyware or a virus, etc., you should assign a **Malicious** flag. Pop-ups that do *not* come back are *not* malicious.

Source: http://vizualbod.com/f/spam-guidelines.htm - 12/03/2008

Part 2: Using EWOQ

1. Introduction

Welcome to EWOQ !

EWOQ is the evaluation system you will use as a rater. You will acquire Tasks and rate them based on the guidelines given to you.

For URL rating, a Task consists of a pair: a **query** and a **URL**. As you work in the EWOQ interface, you will acquire Tasks as you need them and submit your ratings as you complete them.

2. Accessing the EWOQ Rating Interface

There are two different ways to access the EWOQ URL rating interface:

1) Rater Hub: Click on the "Start Rating Now" link in the upper left corner of the Rater Hub homepage. This link will take you to your Rating Home.

2) Go to this link - https://www.google.com/evaluation/search/rating/home

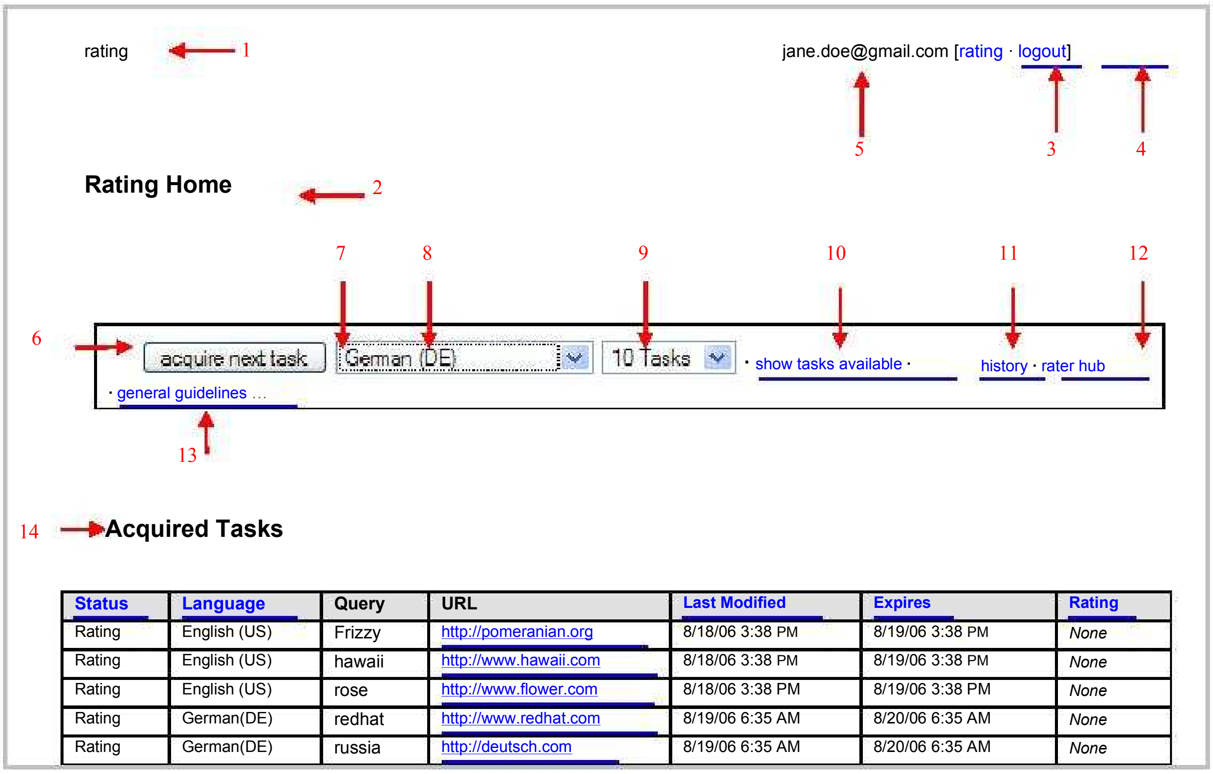
You will supply your GMAIL user ID and password for authentication.

3. Rating

In general, rating a Task involves the following steps:

1.	Acquiring Tasks	(See Section 4)
2.	Starting to rate	(See Section 6)
3.	Submitting your initial rating	(See Section 6)
4.	Re-rating unresolved Tasks	(See Section 7)
5.	Commenting	(See Section 9)

4. Rating Home Screenshot



rating

jane.doe@gmail.com [rating · logout]

Rating Home

Acquired Tasks

Status	Language	Query	URL	Last Modified	Expires	Rating
Rating	English (US)	Frizzy	http://pomeranian.org	8/18/06 3:38 PM	8/19/06 3:38 PM	*None*
Rating	English (US)	hawaii	http://www.hawaii.com	8/18/06 3:38 PM	8/19/06 3:38 PM	*None*
Rating	English (US)	rose	http://www.flower.com	8/18/06 3:38 PM	8/19/06 3:38 PM	*None*
Rating	German(DE)	redhat	http://www.redhat.com	8/19/06 6:35 AM	8/20/06 6:35 AM	*None*
Rating	German(DE)	russia	http://deutsch.com	8/19/06 6:35 AM	8/20/06 6:35 AM	*None*

The red numbers represent the following:

1. **rating**
   This text shows what type of Task you are working on. In this case, the Task type is “rating”.

2. **Rating Home**
   Shows you where you are inside the EWOQ system. This is where you acquire Tasks and where you can see your Task list.

3. **rating**
   Click on this link to go back to the **Rating Home**.

4. **logout**
   Click on this link to end your EWOQ session. Please LOGOUT before ending your EWOQ session.

5. **jane.doe@gmail.com**
   Your GMAIL account.

6. **acquire next Task**
   Click this button to acquire Tasks.

7. **language drop-down menu**
   You can select a language from this drop-down menu.

8. **Local Extension**
   Note that most languages will have a location extension. For instance, Spanish (**MX**) stands for Spanish (**Mexico**) and Portuguese (**BR**) stands for Portuguese (**Brazil**), respectively.

Source: http://vizualbod.com/f/spam-guidelines.htm - 12/03/2008

9. Task

From this drop-down menu you can specify “1 Task”, “5 Tasks”, “10 Tasks” or “20 Tasks”. By default, you will acquire 10 Tasks at a time. EWOQ allows you to have up to 20 Tasks with the status “Rating” in your Task list at any time (this includes new Tasks just acquired and saved drafts of Tasks not yet submitted).

10. show tasks available

Click on this link to display all available Tasks in different language(s) you are assigned to. See Section 5.

11. history

This report shows history information for a chosen period.

12. rater hub

This is the primary resource page which supports the quality rating program. This page contains FAQs, News & Updates, Helpful Suggestions, etc. Please use your Gmail ID and password to gain access to the Rater Hub.

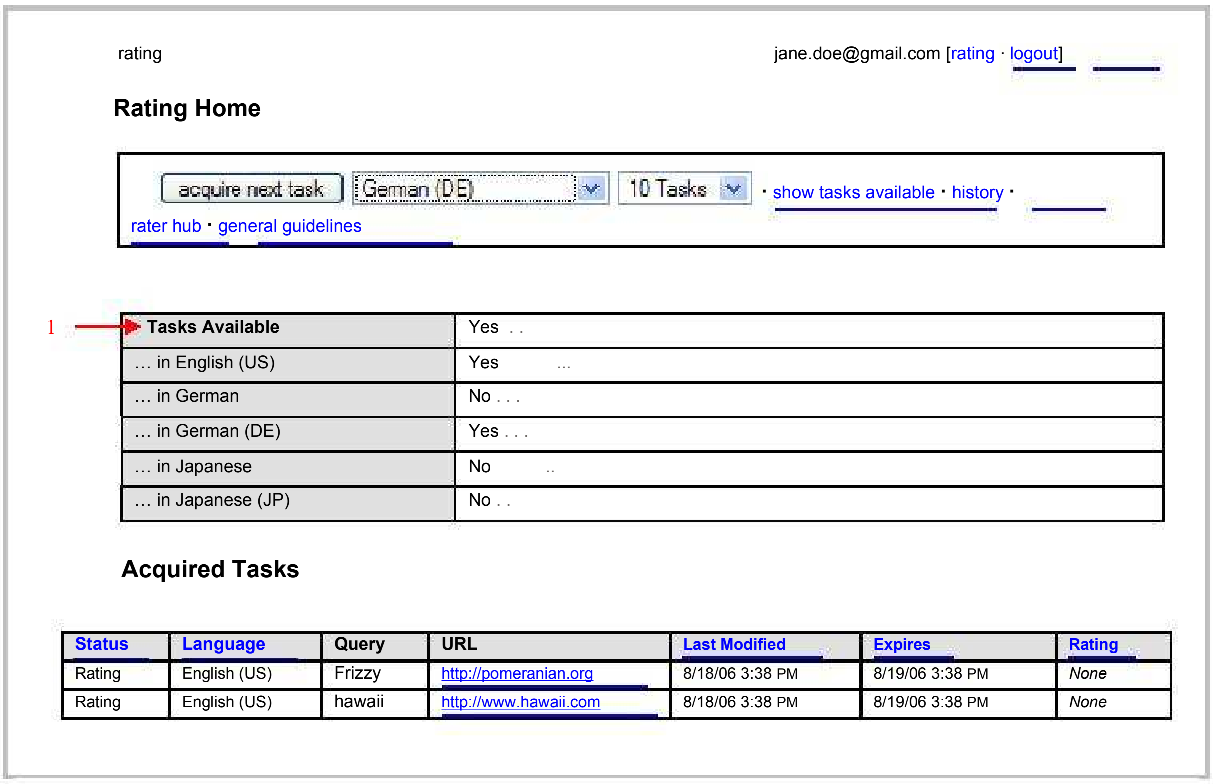
13. general guidelines

Click on this link to read the “General Guidelines”.

14. Acquired Tasks

Shows whether Tasks have been acquired. When no Tasks have been acquired, you will see “No Tasks acquired…”

5. Rating Home Screenshot - after clicking on the show Tasks available link


rating
jane.doe@gmail.com [rating · logout]
acquire next task
German (DE)
10 Tasks
· show tasks available · history ·
rater hub · general guidelines
1

Rating Home

Tasks Available	Yes . .
… in English (US)	Yes ...
… in German	No . . .
… in German (DE)	Yes . . .
… in Japanese	No ..
… in Japanese (JP)	No . .

Acquired Tasks

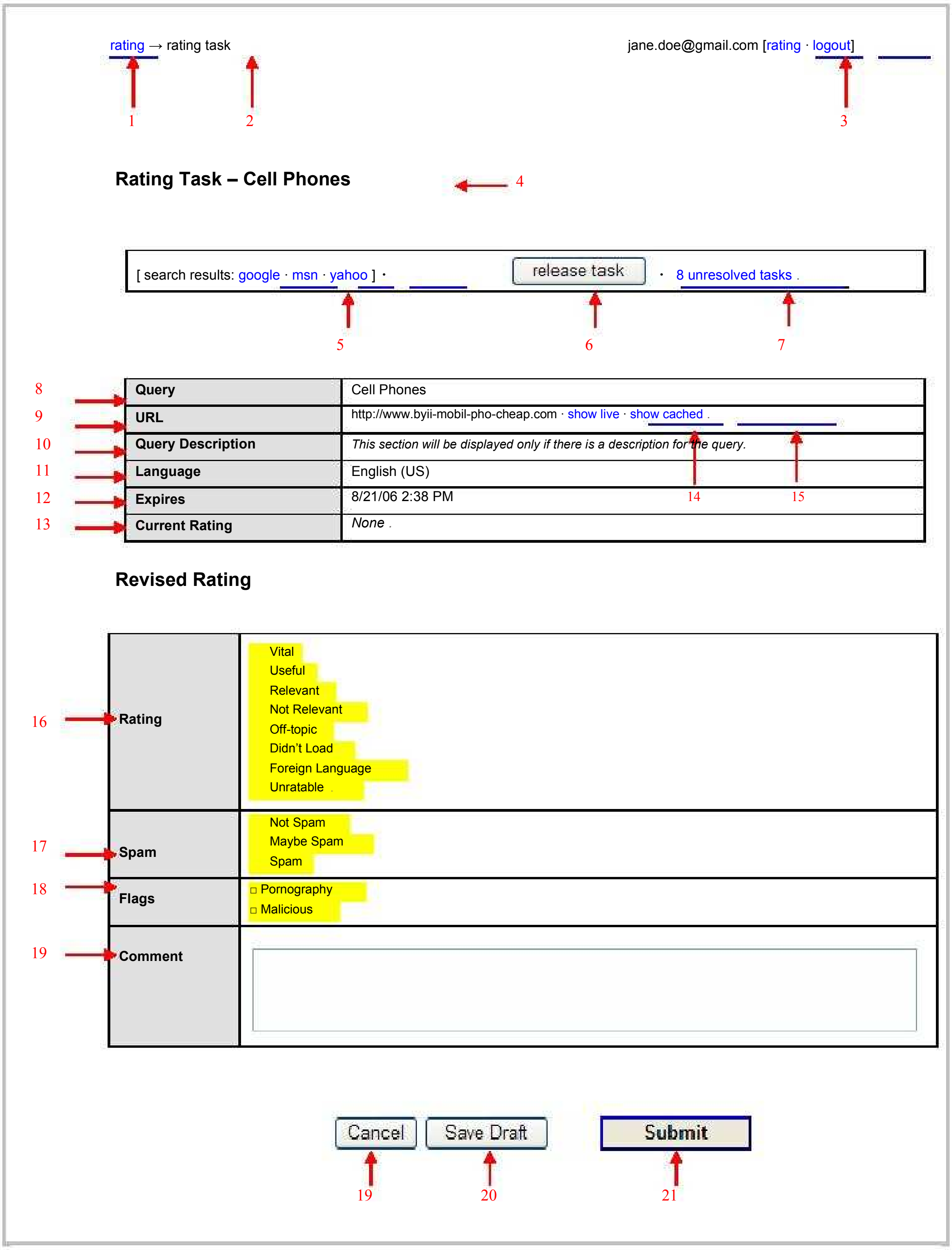
Status	Language	Query	URL	Last Modified	Expires	Rating
Rating	English (US)	Frizzy	http://pomeranian.org	8/18/06 3:38 PM	8/19/06 3:38 PM	*None*
Rating	English (US)	hawaii	http://www.hawaii.com	8/18/06 3:38 PM	8/19/06 3:38 PM	*None*

The red number represents the following:

1. Tasks Available

Click on the show task available link to view this table. The table shows whether Tasks are *currently* available for your language(s).

6. Rating Task Screenshot



rating → rating task

jane.doe@gmail.com [rating · logout]

Rating Task – Cell Phones

[search results: google · msn · yahoo] · release task · 8 unresolved tasks

Query	Cell Phones
URL	http://www.byii-mobil-pho-cheap.com · show live · show cached
Query Description	*This section will be displayed only if there is a description for the query.*
Language	English (US)
Expires	8/21/06 2:38 PM
Current Rating	*None*

Revised Rating

Rating	Vital Useful Relevant Not Relevant Off-topic Didn't Load Foreign Language Unratable
Spam	Not Spam Maybe Spam Spam
Flags	□ Pornography □ Malicious
Comment	

Source: http://vizualbod.com/f/spam-guidelines.htm - 12/03/2008

The red numbers represent the following:

1&3. rating

Returns to **Rating Home** where you acquire Tasks and where your complete Task list is visible.

2. rating → rating task

Shows your location in the EWOQ system; in our screenshot, the display shows the path from **Rating Home** (indicated here by “rating”) to the current **Rating Task** page.

4. Rating Task – [query]

The heading of the **Rating Task** page displays the query.
In our example, the text **Rating Task – Cell Phones**, indicates that the query is **[Cell Phones]**.

5. search results

These links are used for web coverage research using common search engines. Clicking on these links automatically gives the results of the query via the search engine you selected in a separate browser window. Individual settings on browsers and/or toolbars may affect whether the search engine results open in a window in front of the EWOQ interface window, behind the interface window, or not at all if pop-ups are blocked.

6. release task

Clicking on this link allows you to remove the Task from your Task list. To ensure you indeed mean to give up a Task, a dialogue box will appear before the Task is released. This is what releasing the Task accomplishes:

a. The released Task will not be considered part of your workflow.
b. The Task will return to the pool of Tasks, to be reassigned to other raters (and possibly reassigned to you) via a randomized process based on availability and priority.

IMPORTANT: The difference between “Release” and “Unratable” …

Option	**Use this option when:**	**Can the Task (same query and URL pair) come back ?**
Release Task	I don’t want to rate the Task now but maybe LATER. .	Yes .
Unratable *	I don’t EVER want to rate this Task. .	No .

* The “Unratable” category is also described in the “General Guidelines”.

7. unresolved tasks

Each Task will be acquired and rated by a group of raters, each working independently. When all raters for a particular Task have submitted their ratings, the ratings for this Task will be analyzed. If the raters disagree with one another by a wide margin, the Task will be returned to the raters involved for re-rating. It will reappear in your Task list on the Rating Home page with the status “Unresolved” and will be highlighted in yellow to catch your attention. Clicking on this link takes you to **Rating Home** where you can see all Tasks.

8. Query

Make sure you understand the query. Please research the query if you are unsure. If you are still unsure after research, assign “Unratable”.

9. URL

The URL paired with the Query. For rating Tasks, you need to check the URL for each Task to determine the rating, the appropriate Spam label, and any applicable flags.

10. Query Description

This field is only present if there is a description for the query. Currently only a minority of queries carry a description. The query descriptions are entered by Administrators. These descriptions may advise you that the query has been known to bring up a particular type of a result, and offer tips on how to rate this type of result. Some descriptions tell you which interpretation of the query should have the most weight. You may not agree with a query description. If so, be sure to make a comment explaining why you disagree. Rate URLs using common sense, your knowledge, and understanding.

11. Language

Shows you in which Language you are currently working. The screenshot (page 5) displays a Task for English (US) .

12. Expires

Shows when a Task expires. In most cases, a Task will be auto-released and disappear from your Rating Task home if it is not submitted within 24 hours from the time you acquire the Task. Some Tasks will disappear more quickly.

13. Current Rating

Shows the current status of a Task being rated. Until you choose a rating, it will say "*None*".

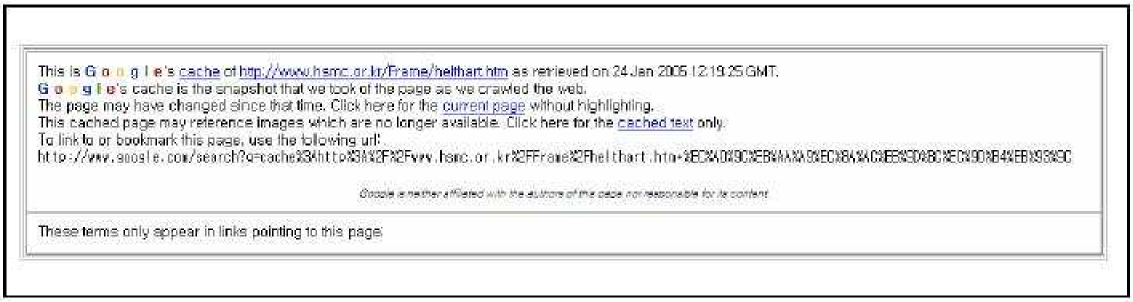
14. show live

Click on this link to rate the page.

15. show cached

Click on this link anytime the live page does not load. The cached page may or may not load. Sometimes, the cached page is blank and you will only see the following information on top with no content below:

Example 1



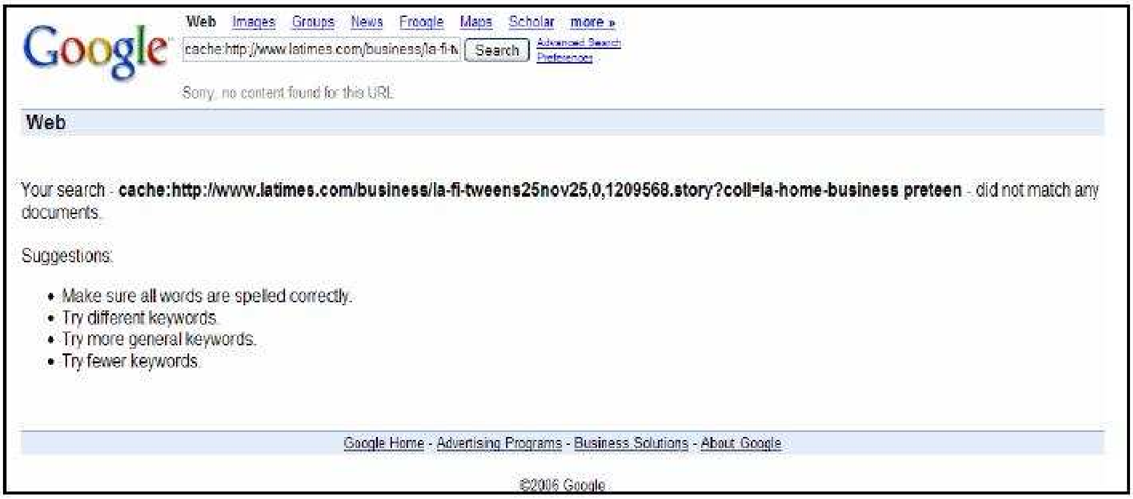
This is Google's cache of http://www.hsmc.or.kr/Frame/helthart.htm as retrieved on 24 Jan 2006 12:19:25 GMT.
Google's cache is the snapshot that we took of the page as we crawled the web.
The page may have changed since that time. Click here for the current page without highlighting.
This cached page may reference images which are no longer available. Click here for the cached text only.
To link to or bookmark this page, use the following url:
http://www.google.com/search?q=cache%3Ahttp%3A%2F%2Fwww.hsmc.or.kr%2FFrame%2Fhelthart.htm+%EC%A0%9C%EB%AA%A9%EC%8A%A4%EB%9D%BC%EC%9D%B4%EB%93%9C

Google is neither affiliated with the authors of this page nor responsible for its content.

These terms only appear in links pointing to this page:

Sometimes there is no cached page. Then will you see a page that looks like this:

Example 2



Google
Web
Images
Groups
News
Froogle
Maps
Scholar
more »
cache:http://www.latimes.com/business/la-fi-t
Search
Advanced Search
Preferences

Sorry, no content found for this URL

Web

Your search - **cache:http://www.latimes.com/business/la-fi-tweens25nov25,0,1209568.story?coll=la-home-business preteen** - did not match any documents.

Suggestions:

- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.

Google Home - Advertising Programs - Business Solutions - About Google
©2006 Google

IMPORTANT: When to rate **live** vs. **cache**:

If live page loads .	Rate the live page .
If live doesn't load, but cache loads .	Rate the cache page .
If neither cache nor live page loads .	Rating is Didn't Load .

16. Rating

Please refer to the "Rating Guidelines" for specific information on each rating category.

17. Spam

Assign one of the three Spam labels to pages that load and can be rated.

18. Flags

Check one or more flags when appropriate.

19. Comment

New raters are REQUIRED to comment on every Task in the initial stage for the first three weeks. After that, commenting is MANDATORY only when you assign a "Spam" or "Maybe Spam" label, and/or a "Malicious" flag.

20. Cancel

You may select "Cancel" to retain a Task without saving any information. Choosing this option will take you back to Rating home and leave the Task in your Task list with the status "Rating".

21. Save Draft

You may revisit a Task at a later time by selecting "Save Draft" to retain your rating and comments. Choosing this option will leave the Task in your Task list with the status "Rating". Please be aware that currently you will not be able to see your comments when you revisit the Task; however, your comments are saved and will be visible during the resolving stage. Even though you cannot see your saved comments, you may add more comments when you access the Rating Task page for the Task at a later time.

22. Submit

You will submit your rating to finalize your work on a Task. *Submitting a Task will remove it from your Task list. You will not be able to revisit or revise the Task*. Submitting a Task will bring you back to Rating Home. You can acquire new Task(s) to rate or work on one of the "Unresolved" Tasks in your Task list.

7. Resolving Tasks (Re-rating Unresolved Tasks) / Moderators

Every Task will be acquired and rated by a group of raters, each working independently. If the raters disagree with one another by a wide margin, the Task will be returned to the raters involved for re-rating in the "resolving" stage. It will reappear in your Task list on the Rating Home page with the status "Unresolved" and will be highlighted in yellow to catch your attention.

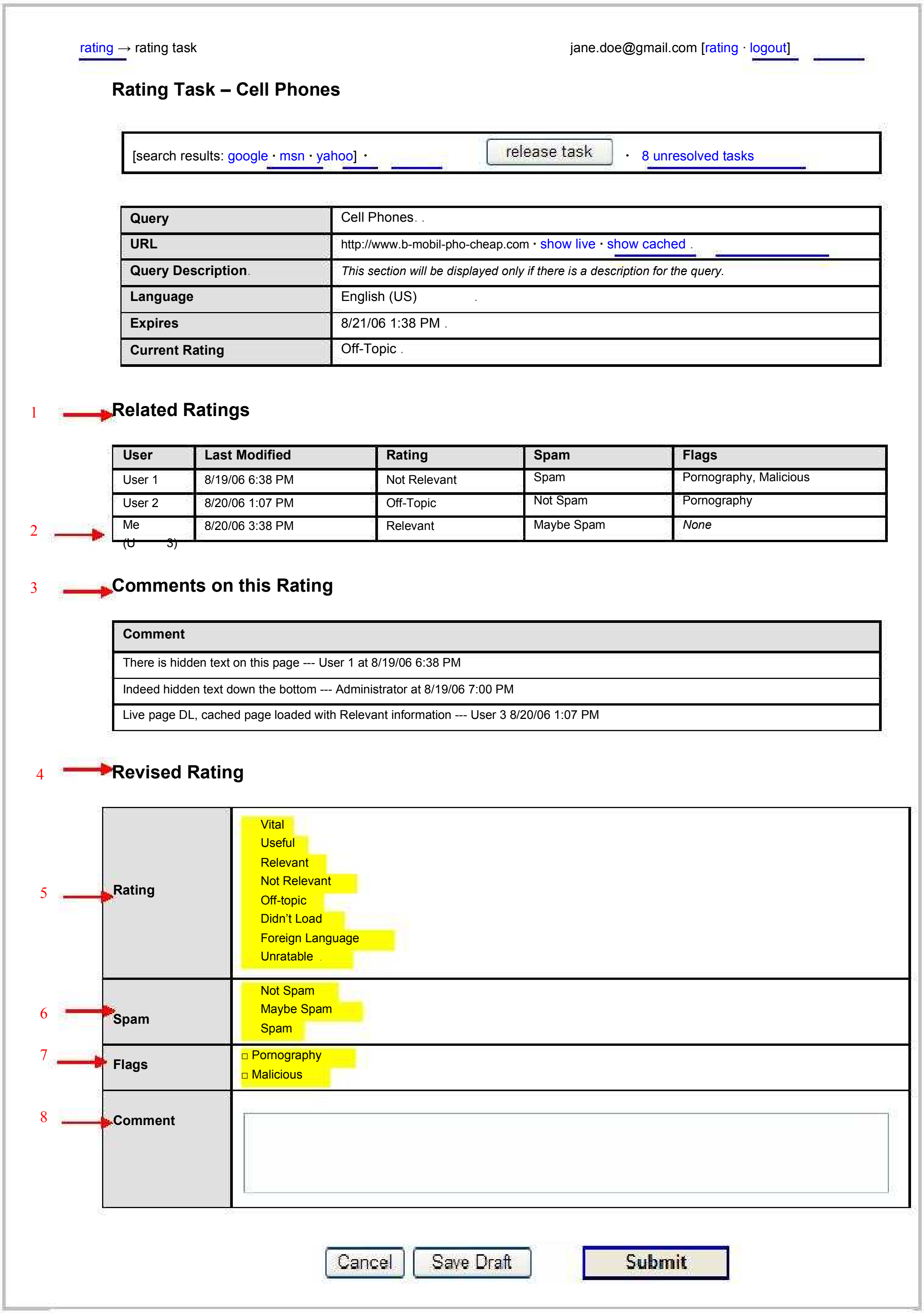
In addition, each time an action has been taken on the "Unresolved" Task by someone other than you, the Task will remain highlighted, but will also be shown in **bold** text. The actions that will cause this to happen are rating changes made by other Raters and/or commenting by Raters, Admins, or Moderators. This is analogous to how unviewed messages appear in bold text in an e-mail inbox. When you see that a Task is in the "Unresolved" state, or that a Task now appears in bold text, please click on it to participate in the resolving process as soon as it is convenient for you to do so.

Moderators

For some unresolved Tasks, you may see comments written by a Moderator. Please pay attention to these comments just as you would comments from an Administrator. The Moderator helps resolve Tasks and contributes to discussions by:

- monitoring Tasks
- highlighting Rater comments
- leaving comments and helpful tips.
-

Source: http://vizualbod.com/f/spam-guidelines.htm - 12/03/2008



rating → rating task

jane.doe@gmail.com [rating · logout]

Rating Task – Cell Phones

[search results: google · msn · yahoo] · · 8 unresolved tasks

Query	Cell Phones. .
URL	http://www.b-mobil-pho-cheap.com · show live · show cached .
Query Description.	*This section will be displayed only if there is a description for the query.*
Language	English (US) .
Expires	8/21/06 1:38 PM .
Current Rating	Off-Topic .

Related Ratings

User	Last Modified	Rating	Spam	Flags
User 1	8/19/06 6:38 PM	Not Relevant	Spam	Pornography, Malicious
User 2	8/20/06 1:07 PM	Off-Topic	Not Spam	Pornography
Me (U 3)	8/20/06 3:38 PM	Relevant	Maybe Spam	*None*

Comments on this Rating

Comment
There is hidden text on this page --- User 1 at 8/19/06 6:38 PM
Indeed hidden text down the bottom --- Administrator at 8/19/06 7:00 PM
Live page DL, cached page loaded with Relevant information --- User 3 8/20/06 1:07 PM

Revised Rating

Rating	Vital Useful Relevant Not Relevant Off-topic Didn't Load Foreign Language Unratable .
Spam	Not Spam Maybe Spam Spam
Flags	□ Pornography □ Malicious
Comment	

Source: http://vizualbod.com/f/spam-guidelines.htm - 12/03/2008

The red numbers represent the following:

1. **Related Ratings**
   Shows the ratings submitted by other raters with a “Last Modified” timestamp. Everyone participating in a Task will stay anonymous. In fact, all raters are addressed by “User” plus a number. Administrators will be shown as Administrator instead of User. Moderators will be shown as Moderator plus a number.

2. **Me (User 3)**
   You will be able to see your initial merit rating with the timestamp. In this example, the rater is identified as User 3.

3. **Comments on this Rating**
   Your initial comments, *if any*, for this Task. As you enter more comments in the future, the comments will be posted in this box. Most recent comments appear on the bottom of the page.

Example 1: User / Moderator

Comment
VITAL– www.wine.com --- User 1 at 8/19/06 6:38 AM .
Can generic subjects have Vital results ? --- Moderator 6 at 8/20/06 2:31 PM .

Example 2: Users / Administrator

Comment
There is hidden text on this page --- User 1 at 8/19/06 6:38 PM .
Indeed hidden text down the bottom --- Administrator at 8/19/06 7:00 PM .
Live page DL, cached page loaded with Relevant information --- User 2 8/20/06 1:07 PM .

Example 3: Users / Moderator / Administrator

Comment
Sneaky redirect to www.sdasdfasde-asdf-zzzz.com --- User 1 at 8/20/06 6:38 PM .
Live DL but cache loads fine with no redirect, therefore DL --- User 3 at 8/20/06 7:00 PM .
Please refer to guidelines for more information on Live vs. Cache and resolve disagreements as soon as possible --- Moderator 8 at 8/20/06 7:30PM
Also check to see if there is any hidden text --- Administrator 8/21/06 1:08 PM .
Sneaky redirect, keyword stuffing and hidden text. Changing from DL to OT/Spam --- User 3 at 8/21/06 2:37 PM

4. **Revised Rating**
   In the box directly below you have the option to choose a rating, select a Spam label and flags, *if any*, and post comments.

5. **Rating**
   For “Unresolved” Tasks, you may either keep the same initial rating or change your rating. Please read comments made by other Users and pay attention to comments left by Moderators and Administrators.

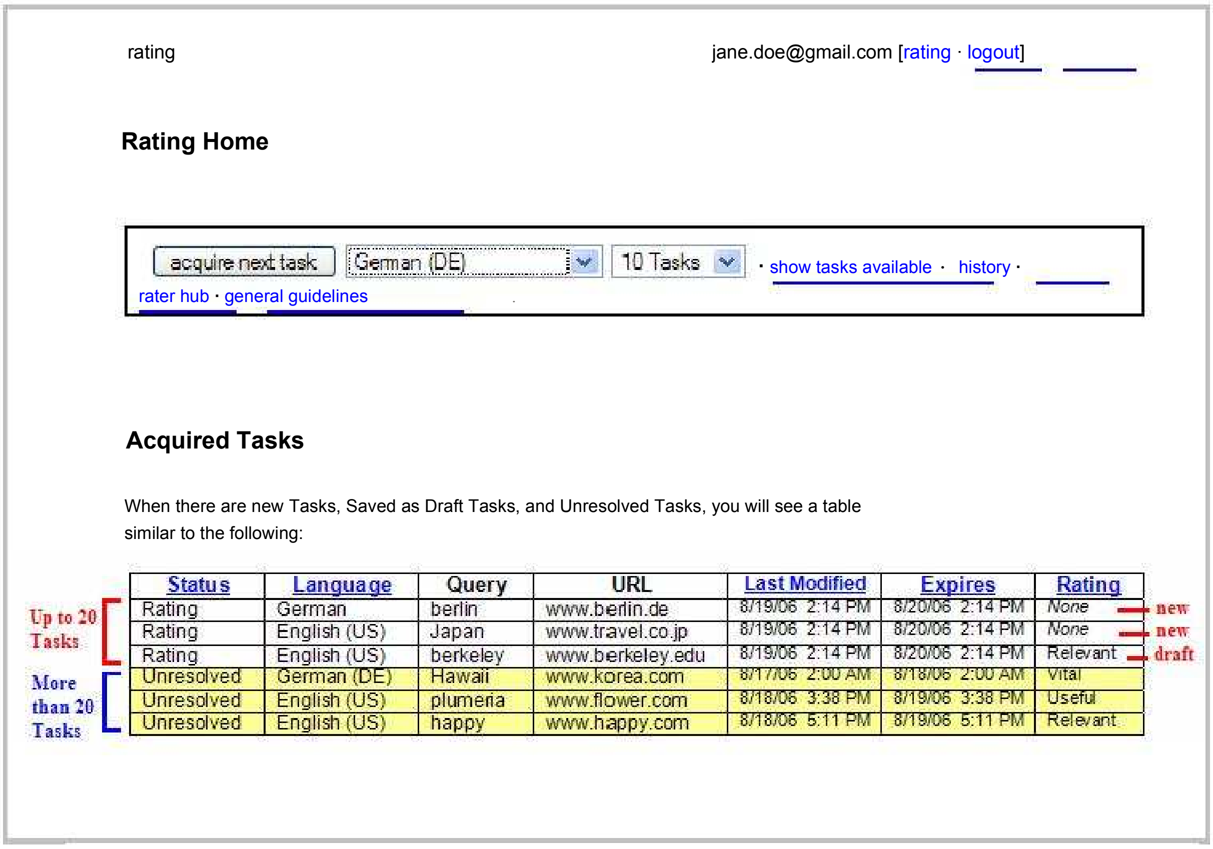
6. **Spam**
   Assign one of the three Spam labels to pages that load and can be rated.

7. **Flags**
   Check one or more flags when appropriate.

8. **Comment**
   When you visit “Unresolved” Tasks, you must leave a comment (in contrast to the initial stage where commenting is mandatory only for “Spam”, “Maybe Spam”, and “Malicious”.) Even if you do not change your initial merit rating, please do leave a brief comment so that everyone knows where you stand. Please read Section 9 on commenting etiquette.

8. Managing Your Task List

The Rating Home page displays your current list of Tasks. You may visit the list at any time by selecting the "rating" link in either of the upper corners of any page in EWOQ.



rating jane.doe@gmail.com [rating · logout]

Rating Home

acquire next task
German (DE)
10 Tasks
· show tasks available · history ·
rater hub · general guidelines

Acquired Tasks

When there are new Tasks, Saved as Draft Tasks, and Unresolved Tasks, you will see a table similar to the following:

Status	Language	Query	URL	Last Modified	Expires	Rating
Rating	German	berlin	www.berlin.de	8/19/06 2:14 PM	8/20/06 2:14 PM	*None*
Rating	English (US)	Japan	www.travel.co.jp	8/19/06 2:14 PM	8/20/06 2:14 PM	*None*
Rating	English (US)	berkeley	www.berkeley.edu	8/19/06 2:14 PM	8/20/06 2:14 PM	Relevant
Unresolved	German (DE)	Hawaii	www.korea.com	8/17/06 2:00 AM	8/18/06 2:00 AM	Vital
Unresolved	English (US)	plumeria	www.flower.com	8/18/06 3:38 PM	8/19/06 3:38 PM	Useful
Unresolved	English (US)	happy	www.happy.com	8/18/06 5:11 PM	8/19/06 5:11 PM	Relevant

***Note**: **Status, Language, Last Modified, Expires**, and **Rating** columns are sortable.*

Prioritization of Tasks in your queue - **Acquired Tasks** (**Status** column)

First Priority "Unresolved"

Second Priority "Rating"

Occasionally, the system runs out of Tasks. Rest assured: new work will come down the pipeline shortly! When the system is out of Tasks, you will see one of these messages:

"No available Tasks were found. Please work on your existing Tasks." Please work on Tasks that are already in your **Acquired Tasks** queue.

"No Tasks of the specified type are available. Please try another." If you are assigned to more than one language, please try different language(s).

9. Commenting Etiquette

The following are guidelines for effective communication during the resolving process in EWOQ.

1. It is important to share relevant background information (reasons, explanations, etc.) when stating your opinion. Indicate your source of information whenever possible. If you come across an important website in your research, please give its full URL.

2. Please do not use abbreviations.
   Exception: To save space and time, the following abbreviations for ratings and flags should be used:

-V	(Vital)	- DL	(Didn't Load)
- Usf	(Useful)	- FL	(Foreign Language)
- Rel	(Relevant)	- MAL	(Malicious)
- NR	(Not Relevant)	- PPC	(pay-per-click)
- OT	(Off Topic)		

Please refrain from using message board lingo (IMO, FWIW, AFAIK, etc.).

3. Write concisely. Do not make unnecessary comments such as "Oh, I see your point" or "Sorry, I missed that". But do write enough to explain yourself clearly to other raters who might not have your background or expertise.

4. Sometimes the most efficient way to make your point is to quote guidelines or other rating information from the Rater Hub. Please be very specific about how the information you quote relates to the situation at hand. When quoting from the Guidelines, please include the version number and page number.

5. When commenting on a query, describe your interpretation of user intent. This is very important for ambiguous or poorly phrased queries. You may include whether you believe the query is navigational, informational, or transactional (to buy, to download, etc.); broad or specific. If you disagree with the Query Description you see on the EWOQ interface, please be explicit about that as well.

6. State your reason for assigning "Spam", "Maybe Spam", and "Malicious" flags. For example,
   - Sneaky redirect to eBay
   - Amazon thin affiliate
   - Keyword stuffing
   - Wikipedia content plus ads
   - DMOZ content plus ads
   - Copied text from (for instance, Wikipedia) plus ads
   - Parked domain
   - Hidden text

7. Brief comments to confirm your rating in the resolving stage are always appreciated:
   - "Still DL for me."
   - "Confirming Usf: it's the best result I could find."

Part 3: Rating Examples

1. Introduction

For select frequent types of queries, we will provide examples, along with suggested ratings based on the Rating Guidelines. Some of the examples are prefaced by definitions or notes on possible interpretations, which are the result of the type of background research that Raters should routinely perform when encountering a new query.

2. Named Entities

This category primarily covers the following types of named entities:

- People (celebrities, public figures, ordinary people, etc.)
- Geographic locations
- Companies, products, and brand names
- Organizations and other institutions
- Titles of books, shows, musical pieces, etc.
- Events

Query	[Leon Harris], English (US)
Note	• *There is a Leon Harris who used to be an anchor for CNN's Live Today and later joined ABC7.* • *There are potentially many other different people named Leon Harris in the world.*
Vital	• ABC Homepage for Leon Harris or Leon Harris' personal homepage.
Useful	• Quality pages with biographical or good general information on ABC's Leon Harris. • The homepage for someone named "Leon Harris", other than the ABC reporter, whose homepage would be of interest to a significant number of people.
Relevant	• Any page with at least a paragraph about ABC's Leon Harris. • News articles about Leon Harris during his time as a CNN anchor. • The homepage for someone named "Leon Harris" other than the ABC reporter, whose homepage would be of interest to a small number of people. • Any quality pages about anyone named "Leon Harris".
Not Relevant	• Pages that mention Leon Harris but are not about him. This includes news articles that seem outdated. • Lower quality pages about anyone named "Leon Harris".
Off-Topic	• Pages by or about people with "Leon" or "Harris" as part of their name. An example would be "Leon Shapiro in Harris County" or "Bob Leon Harris".

Source: http://vizualbod.com/f/spam-guidelines.htm - 12/03/2008

Query	**[Nicole Kidman] , English (US)**
Note	*Nicole Kidman is known as one of Hollywood's top movie stars.*
Vital	• Nicole Kidman's official page. Be aware that other sites may claim to be official.
Useful	• Pages that are comprehensive resources for Nicole Kidman; a comprehensive resource would include her biography, filmography, pictures, etc. The page might even be a personal fan page.
Relevant	• News articles of at least one paragraph with timely and informative material about Nicole Kidman.
Not Relevant	• Pages containing little information about Nicole Kidman.
Off-Topic	*Note: Well-known actresses and personalities are often exploited for porn/spam.* • Off-Topic + Spam – http://www.nicolekidman.org.

Query	**[Chicago] , English (US)**
Note	*Chicago is a big city in America.*
Vital	• The official homepage for the city of Chicago http://egov.cityofchicago.org/city/webportal/home.do .
Useful	• The homepage for the main regional newspaper, Chicago Tribune http://www.chicagotribune.com/. • High quality pages about Chicago: history, climate, travel, etc. • An excellent personal collection of information can qualify for this rating.
Relevant	• The homepages of large, prominent entities associated with the word 'Chicago' in the minds of many people. The Chicago Bulls and The University of Chicago are examples. • Homepages of relatively important universities or businesses in the Chicago area, or newspapers that cover the Chicago area but are not the main 'paper of the city' can be **Relevant** or **Not Relevant**.
Not Relevant	• Local weather forecasts for Chicago • Pages for businesses or other organizations located in Chicago, other than the well-known sports teams, universities, businesses, newspapers, or other entities. • Websites about Chicago in general.
Off-Topic	• Pages that merely mention any of the organizations whose homepages are classified as **Relevant** or **Not Relevant** above (e.g. a journal article to which a professor at Northwestern University in Chicago has contributed).

Note: Major cosmopolitan cities are preferred targets for spammers, especially hotel affiliates. Such results should be labeled as Spam, even if they have relevance to the query – e.g. a hotel affiliate page with a list of Chicago hotels may be **Relevant**.

Query	[freeman manufacturing] , English (US)
Note	*An ambiguous query. There are at least two companies, http://www.freemansupply.com/ and http://www.freemanmfg.com/ that can be likely targets of this search.*
Vital	• http://www.freemansupply.com/ the homepage for the company, Freeman Manufacturing and Supply Company that has the words "Freeman Manufacturing" in its name and is a large, widely known entity.
Useful	• Subpages of the Freeman Manufacturing and Supply Company site, provided the subpages are important: e.g. they give company info, or info about some of main lines of products, such as Liquid Tooling Materials. • Websites selling or reviewing Freeman Manufacturing and Supply Company products. • http://www.freemanmfg.com/ the homepage of the orthotics/prosthetics company that does *not* have the word "manufacturing" in the title (but has this word abbreviated in the domain name) and is a smaller entity.
Relevant	• Websites selling or reviewing Freeman Orthotics/Prosthetics products. • Subpages of Freeman Orthotics/Prosthetics, provided the subpages are important: e.g. they give company info, or info about some of main lines of products. • The homepage for other companies called "Freeman Manufacturing" that may exist. • Pages that provide a link to either site with some descriptive or contact information http://www.globalspec.com/Supplier/profile?vid=122697
Not Relevant	• Unimportant subpages on either of the above mentioned sites. • Pages that provide a link to one of the sites without giving any descriptive information, so that the value of the page is entirely in the link it provides.
Off-Topic	• Pages with the words "Freeman" and "Manufacturing" but not about any entity called "Freeman Manufacturing".

Query	[A O Smith] , English (US)
Note	*A.O. Smith is a company that makes electric motors, water heaters & storage tanks.*
Vital	• Corporate Homepage for A.O. Smith http://www.aosmith.com/
Useful	• A.O. Smith Home Pages other than the main corporate page listed above, such as http://www.aosmithmotors.com/ and http://www.hotwater.com/ • Important subpages on the A.O. Smith website. • High quality websites about the A.O. Smith Corporation, or sites that sell, distribute or review A. O. Smith products.
Relevant	• A small fact sheet by a trusted source.
Not Relevant	• Unimportant subpages on the A.O. Smith website
Off-Topic	• Websites that have the word "Smith" and the letters "A" & "O", but are not about "A.O. Smith".

Source: http://vizualbod.com/f/spam-guidelines.htm - 12/03/2008

Query	[xw9400] , English (US)
Note	*xw9400 is a workstation (computer) made by Hewlett Packard.*
Vital	• HP's homepage for the xw9400 http://h10010.www1.hp.com/wwpc/us/en/sm/WF05a/12454-296719-296721-307907-296721-3211286.html
Useful	• High quality sites that sell, review, or provide comprehensive information on the xw9400.
Relevant	• Newsgroup postings about the xw9400 can be **Relevant**, but do not have to be. • If this is a new product, news releases about the workstation. • Price comparison listings for the product.
Not Relevant	• Pages about HP computers in general or HP computers other than the xw9400 • Pages that mention the HP xw9400 but are not focused on it.
Off-Topic	• Pages on other brands of computers.

Query	[people search] , English (US)
Note	*This is an example of a query that can be interpreted as generic (give me sites that help look for people) or a named entity (give me the homepage of the company by that name).*
Vital	• Homepage for Yahoo! People Search: http://people.yahoo.com/ , a widely known resource.
Useful	• High quality "people finder" sites that allow you to search for people. Must search a broad geographic area, however, not just within a company, university or organization. The search functionality must work, otherwise **Not Relevant**.
Relevant	• Services that provide background checks on individuals for a fee http://peoplesearch.com/business/ (note that despite the match between the query and the URL this result is only **Relevant**, as would be pages of other companies engaged in the same business).
Not Relevant	• People finder services that search within a geographic area outside the U.S. For instance, in Australia only. • "People finder" search boxes on personal pages.

3. Informational Queries

This section provides examples on a wide range of informational queries. General categories include:

- Nature, Science, Medicine, and Health
- Culture
- Consumer Information and Practical Advice
- Technical (these queries frequently are on the border between Informational and either Navigational or Transactional)

Source: http://vizualbod.com/f/spam-guidelines.htm - 12/03/2008

Query	[retina and laser surgery] , English (US)
Vital	• None possible.
Useful	• Pages from high quality sources providing information on laser surgery for the retina. • High quality FAQ's with extensive information. • Newsgroups which are focused on the subject and provide helpful information.
Relevant	• Individual laser surgery practitioner websites that provide relevant information.
Not Relevant	• Newsgroups discussing the topic in very general terms.
Off-Topic	• Sites about cosmetic laser surgery (not surgery on eyes); sites about eye surgery without the use of lasers.

Query	[what can I do with coffee grounds], English (US)
Vital	• None possible.
Useful	• Any page giving quality advice on uses for coffee grounds (deodorizer, fertilizer, etc.). • FAQ's and newsgroups, if they provide good resources for uses of coffee grounds.
Relevant	• A page that provides a single use for coffee grounds.
Not Relevant	• Pages that talk about coffee grounds, but give no advice on what to do with them.
Off-Topic	• Businesses or web sites called "Coffee Grounds" but that do not give advice on uses for them.

Query	[HTML lessons], English (US)
Note	*HTML stands for HyperText Markup Language, which is the predominant markup language for the creation of web pages..*
Useful	• Pages providing lessons or step-by-step instructions or tutorials for learning HTML. The pages should be good general resources for learning HTML, for example, they should not focus on just one advanced aspect of HTML.
Relevant	• General resources on HTML. • Pages that focus on just one aspect of HTML.
Not Relevant	• Pages that discuss HTML, but do not provide lessons.
Off-Topic	• Pages about "XML", but not about "HTML".

Query	[disable javascript ie], English (US)
Note	*"IE" (Internet Explorer) is Microsoft's web browser.*
Vital	• The page from Microsoft's website which tells you how to disable Javascript in IE.
Useful	• Other authoritative websites that help the user disable javascript in IE.
Off-Topic	• Websites limited only to the Netscape Navigator browser.

Source: http://vizualbod.com/f/spam-guidelines.htm - 12/03/2008

Query	**[adobe reader download], English (US)**
Note	*Adobe Reader software allows the viewing and printing of PDF files.*
Vital	• The page on the Adobe site on which to download Adobe Reader. http://www.adobe.com/products/acrobat/readstep2.html
Useful	• A page that thoroughly explains what "adobe reader" is or does. • The Adobe homepage. http://www.adobe.com/
Relevant	• Pages that just discuss Adobe Reader.
Not Relevant	• A page with just a link to the page on the Adobe site at which to download Adobe Reader.
Off-Topic	• A page on which to download the Omea Reader. http://www.jetbrains.com/omea/download/reader.html

4. Targeted Information Queries

This category covers queries that help the user obtain the targeted information exactly. Examples include but are not limited to: downloads, maps, lists, contact information, store hours, etc.

Query	**[map biscayne st south beach], English (US)**
Useful	• Map that shows the South Beach area of Miami Beach, and identifies Biscayne Street.
Relevant	• Map that shows the South Beach area of Miami Beach, but does not identify Biscayne Street.
Not Relevant	• Pages about South Beach, Miami, or Florida, in general.
Off-Topic	• The official site of the Biscayne National Park in Florida. • Map finder page in which you can type “biscayne st, south beach” and get a map of the area.

Query	**[international telephone codes], English (US)**
Useful	• A page that displays a list of the international telephone codes (dialing prefixes) for foreign countries (“39” for Italy, “49” for Germany, etc.). It MUST be a comprehensive list of countries. Drop down lists, hyperlinked letters of the alphabet that lead to alphabetically arranged lists of countries, or search boxes that return the country code for a specified country are fine, as long as the search feature functions.
Relevant	• Pages with international telephone codes, but with only a limited selection of cities in major countries.
Not Relevant	• Descriptions of how to dial internationally, info about phone scams, etc. • Pages that list just a few country codes as part of something much larger (e.g. how to use a phone system, etc.)
Off-Topic	• Pages with U.S. area code lists.

Query	[recipes chicken breast], English (US)
Useful	• Quality pages with many recipes using chicken breast. The recipes should use chicken breast, not other parts of the chicken. • Quality sites with links to many recipes made with chicken breast. • 'Search for recipe' pages on trustworthy recipe websites, such as http://www.epicurious.com/.
Relevant	• Pages that have just one or two chicken breast recipes on them.
Not Relevant	• Pages that predominantly have recipes for whole chickens, or dark meat (legs, thighs). • Pages about recipes in general.

Query	[Louvre visiting hours], English (US)
Vital	The subpage on the site of the Louvre that provides visiting hours information with unique authority http://www.louvre.fr/llv/pratique/horaires.jsp?bmLocale=en
Relevant	• A page about the Louvre from a museum guidebook, with hours mentioned. http://www.europeanmuseumguide.com/museumInfo.php?museumid=115
Not Relevant	• General travel information about Paris with a brief mention of the Louvre without reference to the visiting hours.
Off-Topic	• Web sites discussing France in general or about Paris without references to the Louvre

5. Queries That Ask for a List

Query	[sedona real estate]
Note	*Sedona is a town in northern Arizona*
Useful	• High quality pages providing real estate listings, or a general overview of real estate in Sedona, Arizona. Homepages for individual real estate offices located in Sedona, AZ.
Relevant	• Pages offering some information about Sedona Real Estate
Not Relevant	• General Directory pages about Sedona even if they offer a real estate link.
Off-Topic	• Pages about Sedona, in general.

Query	[London Boutiques]
Useful	• High quality sites with lists of major shops, including maps and store hours. • A page focused on London's women's fashion shops. It might include pictures, addresses, descriptive information, and price ranges. http://www.talkingcities.co.uk/london_pages/shopping_womensfashion.htm
Relevant	• A personal site with recommendations of where to shop in London (must be appropriately specific, i.e. have enough content on boutiques, not warehouses).
Not Relevant	• The homepage of a guide to London that includes a link to shops in London http://www.talkingcities.co.uk/london_pages/london_main.htm. • A page on a site with spotty information on shopping in London
Off-Topic	• Page about boutiques in Liverpool • The homepage of a guide to London that contains no information about shopping.

Part 4: Webspam Guidelines

WHAT IS WEBSPAM ?

Webspam is the term for web pages that are designed by webmasters to trick search engine robots and direct traffic to their websites. In the coming pages, you will learn how to identify some of these techniques. When you observe them being used, you will assign a Spam label to the page.

The Relationship between Ratings and Spam

You have already learned that pages are rated according to their relevance to the query and utility to the user. You would not be able to assign a rating without knowing the query. We say that ratings are **query-dependent**.

Spam labels do not depend on the relevance of the page to the query. Spam labels are **query-independent.** A page should receive a Spam label if it is created using deceptive techniques - no matter what query it is associated with. It is possible for a page to receive a very high rating – even a Vital rating – and also be assigned a Spam label.

How do Spammers make money from the use of Spam?

Spammers make money when visitors click on links on their web pages. Revenue sources are of two general types:

Pay-Per-Click (PPC) ads: Spammers make money each time an ad is clicked. PPC ads appear on many different types of web pages. Sponsored links is another term for ads.

Thin Affiliates: Spammers make money when a transaction is made after the user has clicked through to the merchant's site.

Exceptions: Pages should generally **not** be marked Spam if they provide **added value.** Added value refers to original or other useful content on the page, regardless of whether there are PPC ads. Examples of content that provides added value include:

- Price comparison functionality: Even though the user has to go to another site via the affiliate link to place an order, there is value to have price comparisons right there on the page.
- Product reviews: Pages that provide original reviews offer added value. Items that are commonly reviewed are books, electronics, and hotels.
- Recipes: Pages that provide recipes offer added value.
- Lyrics and quotes: Pages that display lyrics or quotes offer added value.
- Contact information: Pages that provide contact information, especially physical addresses and phone numbers, offer added value.
- Coupon, discount, and promotion codes: Affiliate pages that provide coupon, promotion, or discount codes for the consumer offer added value.

TYPES OF SPAM

This section describes the following types of Spam and provides tips and tools on how to identify them.

- PPC Pages
- Parked Domains
- Thin Affiliates
- Hidden Text and Hidden Links
- JavaScript Redirects
- Keyword Stuffing
- 100% Frame
- Sneaky Redirects

1. PPC Pages

Many web pages are set up for the purpose of collecting pay-per-click (PPC) revenue without providing any or much content of their own. These pages will frequently look like search results, or they may look like a blog or message board (forum) pages. There are many different types of PPC pages:

Pages with PPC Ads only: Some pages contain nothing but PPC ads (or sponsored links).

Fake Directories with PPC Ads: With a fake directory, you will see a list of links that look like search results. However, clicking on a few links reveals that they are just ads disguised as "results".

Example of a fake directory: http://www.favse.com/search.php?q=online+kitchen+design+tool

Fake Blogs with PPC Ads: With a fake blog, you will see an entry that is either nonsensical or copied from another source. The page exists so that the links on the page will be clicked.

Example of a fake blog: http://isinternetbackgammoncom.blogspot.com/

Fake Message Boards with PPC Ads: With a fake message board, you will see "messages", but you will not see responses to the messages. The text in the message may be nonsensical or the "message" may contain PPC links within it. There may also be PPC links on the page. You may actually find entire copied forums that have been scraped from various sources that provide content. The sites may appear to offer comments, registration, and login sections, but when you attempt to use them, they either don't work at all or you land back on the same page.

Scraped or Copied Content with PPC Ads: Scraped or copied content refers to content that has been stolen from another source, either through the use of a piece of software that searches for content containing specific keywords, or through simple copy-and-paste. It also refers to content obtained from sources that allow for distribution and may even encourage re-use, such as Wikipedia and DMOZ. Some of the sources that are routinely scraped or downloaded from by spammers are:

- http://www.wikipedia.org/ : A human-edited online encyclopedia that is freely available for download and re-use.
- http://www.dmoz.org/ : The Open Directory Project, a human-edited directory of the Web also available for download.
- RSS (Really Simple Syndication) and XML (Extensible Markup Language) news feeds: web publishers (such as the BBC, CNN, Usenet, CNet, NYTimes, and others) publish information online that is readily available to users.
- Scraped search results from other companies: Overture.com and Searchfeed.com, among others, provide feeds of PPC search results to most qualifying webmasters.
- Templates: Some sites utilize templates to mass-reproduce web pages automatically. The content is usually scraped from sources that provide such content. You will learn to recognize these templates which usually follow a generic format or pattern.

Please note that the acquisition of content from these sources is not necessarily illegal, nor plagiarism. Webmasters who create copies usually do not claim to be original content creators and may, in fact, assign credit to the originator of the content.

Recognizing Scraped Content

You can copy a snippet of text (a sentence or part of a sentence) and paste it in the search box to see if you can find its source. You will sometimes discover that the text was copied from Wikipedia or one of the other sites mentioned above, or you may find that the text exists on many, many web pages.
You will see various revenue sources (PPC ads) surrounding the content, unlike the original sources (Wikipedia, DMOZ, etc.) that display no ads.
After a while, you will become familiar with the format of Wikipedia pages, particularly the section headings and links provided.
Similarly, you will become familiar with DMOZ pages, which utilize a directory pathway. In addition, these pages offer links to DMOZ that invite you to "submit a site" or "become an editor".

You can do a 'site:' search to look for URL formatting that suggests that a template was used. For example, if the questionable URL is www.might-be-spam.com, you would type "site: http://might-be-spam.com" in the search box to see how many times it appears.
You can look for suspicious "computer-manufactured" grammar.

Example of Wikipedia scraped content with ads: http://www.dgun.org/en/Estonia

Exceptions (Scraped Content that is not Spam)

Lyrics, poems, ringtones (that the user programs rather than downloads), quotes, and proverbs have no central authority. When you see pages with this content, you cannot judge it to have been copied, and the pages should not be assigned a Spam label.
Unfortunately, some content is written specifically for Spam pages and you will not find it on another source. Although you may be convinced that the intent is to deceive, if the content makes sense and appears original, you will not be able to label such pages Spam.

Sometimes the viewing area contains nothing but ads, but there may be scraped content positioned well "below the fold" – on the lower portion of the web page that the user wouldn't see without scrolling down. **The important thing to remember is that if the scraped (copied) content on the page is removed and all that remains is ads, it is Spam.**

2. Parked Domains

A domain name, whose renewal date has passed but which has not yet been dropped from the DNS (domain name system), may be purchased by new owners. Spammers sometimes buy these domains and put their own content on the site. The sites are referred to as parked or expired domains, and their value is in their pre-existing links. Pages that previously linked to the expired domain will now link to the spammer's page.

A typical parked/expired domain may include:

- A list of sponsored links.
- A list of popular categories.
- A list of related categories.

All of the links are paid links. There is no original content on the page.

Recognizing Parked/Expired domains

Look for a domain name (URL) that has nothing to do with the content on the web page.
Check http://www.waybackmachine.org to see the site as it looked previously.
Before long, you will become familiar with the layout of parked/expired domains.

Example of a parked domain: http://www.dasonet.com/todahfzkdk.htm

3. Thin Affiliates

A **thin affiliate** is a page that exists to deliver a visitor to a page on another domain with a different owner. Keywords deliver visitors to the affiliate page, and links on the affiliate page deliver visitors to the second page, which is owned by a real merchant.

This is a revenue-sharing situation in which the thin affiliate is paid a commission by the real merchant for any activity generated on the merchant's site. Usually the activity will be a sales transaction, such as a product purchase or a hotel booking. The thin affiliate site contains text and perhaps images copied from the merchant site. It offers no (or very little) value-added service while earning its commission. The thin affiliate may also earn PPC revenue by providing PPC links on its page.

Recognizing thin affiliates

Clicking on a "More Information" or "Make a Purchase" button takes you to a merchant on a different domain.
Right-clicking on an image on the page with your mouse and looking at "Properties" shows you that the image has a different URL than the URL of the page. This indicates that the image originates from the other merchant's site.

Example of a thin affiliate: http://findmeatune.com/artist-Pink
This is an Amazon thin affiliate. Looking at the properties of the images on the page reveals that they originate from amazon.com.

There is no value added (e.g. reviews, price comparison) on the page, and the value of the page is only in the link to the merchant's site.
You cannot complete a transaction from the thin affiliate's site.

Many large web retailers offer affiliate programs. Some of the most he most common examples are Amazon, eBay, Zappos, and Overstock.

Recognizing true merchants

Features that will help you determine if a website is a true merchant include:

a "view your shopping cart" link that stays on the same site and updates when you add items to it,
a return policy with a physical address,
a shipping charge calculator,
a "wish list" link, or a link to postpone purchase of an item until later,
a way to track FedEx orders,
a user forum,
the ability to register or login,
a gift registry, or
an invitation to become an affiliate **of** that site

Please note the following:

Not all of the above need to be present for a merchant to be considered a true merchant.
Yahoo! Stores are true merchants – they are not thin affiliates.
Some true merchants will take you to another site to complete the transaction due to the fact that they utilize third party cart systems. Such merchants are not thin affiliates.

Not all affiliates are thin

If a page offers some value in addition to its links to the merchant, then it is not a thin affiliate. For example, if the affiliate offers price comparison functionality, or displays product reviews, recipes, lyrics, etc., it is not a thin affiliate, and, therefore, not Spam. Some companies that offer price comparisons or other helpful shopping features in addition to the affiliate link are:

http://www.shopping.com
http://www.pricegrabber.com
http://www.kelkoo.co.uk

4. Hidden Text and Hidden Links

Webmasters add hidden text to lure users to their pages. The hidden text is visible to the search engine robot, but not to the user, who might find it distracting or distasteful.

The text may be completely invisible to the human eye.
The text may be in a very close color to the background on the page so that it is almost invisible and won't be noticed by the human eye.
The text may be formatted in a very, very small font size (e.g., 1-point) so that it won't be noticed by the human eye.
The text may be placed outside the normal viewing area. For example, the webmaster may place a large blank space between the normal viewing area and a "hidden" area all the way at the bottom of the page or far to the right.

Please note that hidden text is not considered to be Spam if there is no intention to trick the search engine. For example, if the webmaster "hides" the date of an update or copyright information either completely or in a very small font size, that would not be considered Spam.

Recognizing hidden text and hidden links

Apply Ctrl-A (the keyboard shortcut for Select All) to the page and then scroll through it. This technique may expose text or links that are hidden from the human eye.

Examples of hidden text:
http://www.bigraf.it/
http://www.h5.dion.ne.jp/~cozmo/
With both of these examples, you should apply Ctrl-A to the page and scroll down on the page.

Be suspicious of large blank areas on the bottom or far right portion of the page, and use the vertical and horizontal scroll bars to see if there is text on the portion of the page outside the main viewing area.
View the source code to see if text exists that is hidden from the user:

If you are using Internet Explorer:	If you are using Firefox:
1. Go to "**View**". 2. Click on "**Source**".	1. Go to "**View**". 2. Click on "**Page Source**".

Spammers commonly employ features of JavaScript to hide text. To disable JavaScript so that you are able to see the hidden text, follow these steps:

If you are using Internet Explorer:	If you are using Firefox:
1. Go to "**Tools**". 2. Click on "**Internet Options**". 3. Click the "**Security**" tab. 4. Click on "**Custom level**". 5. Scroll down to the "**Scripting**" section. To disable JavaScript, make sure "**Disable**" is selected under "**Active scripting**". 6. Click "**OK**".	1. Go to "**Tools**". 2. Click on "**Options**". 3. Click on "**Content**" or "**Web Features**". 4. To disable JavaScript, make sure the "**Enable**" box is not checked. 5. Click "**OK**".

After disabling JavaScript, view both the live and cached versions of the page, because sometimes the hidden text will only be revealed on one page or the other.

Spammers may also use JavaScript to create two versions of their content: one to be viewed and ranked by the search engine, the other to be seen by the user. You can also use the steps outlined above to view the different pages.

Minute text is not always exposed using Ctrl-A. Be suspicious of horizontal lines or bars on the page. Sometimes they contain hidden text. Use the techniques above to check for it.
Some webmasters employ CSS (Cascading Style Sheets) to transform text into minute size or to hide it. Such ploys are not easy to spot, and identifying them by disabling this feature is an advanced technique and is totally optional.

5. JavaScript Redirects

As you have seen above, webmasters sometimes use JavaScript features to hide text. They may also use it to create two versions of their content: one to be viewed and ranked by the search engine, the other to be seen by the user.

Recognizing JavaScript Redirects

If you suspect that a page is Spam and the cached page is available, you should compare it to the live version. A significant difference between the two can be a spam signal. You can also use the steps outlined above to view the different pages. You should observe both the live and the cached pages with JavaScript disabled.

6. Keyword stuffing

Webmasters sometimes load pages with **excessive keywords** that are **related** to the content on the page to draw search engine robots to their web pages. These will appear in the form of word repeats, related words, and misspellings.

Webmasters also sometimes load pages with **irrelevant (off-topic) keywords** (pertaining to topics such as mortgage, cell-phones, gambling, weather…) that are **unrelated** to the content on the page. Again, the intent is to increase traffic to their web pages.

Recognizing keyword stuffing

Keyword stuffing can be found anywhere on the web page**.** In some cases, the keyword stuffing is visible to the human eye, and you will not have to use any special tools to see it. In other cases, it is used in connection with hidden text, in other words, the text that is hidden contains keyword stuffing. When this is the case, you will use the techniques described above to uncover it.

> Examples of hidden text/keyword stuffing/different page uncovered by disabling JavaScript:
> http://equal.smilebo.net/1310nm-is-equal-too.html/
> http://skipper.aalimprincess.com/skipper-key.html

Keyword stuffing in the URL

URLs may also contain keyword stuffing. They are usually created by some type of template and are stuffed with terms that come from the query. They are often formatted with many hyphens (dashes) in them.

> Example of keyword stuffing in the URL:
> http://apply-bankruptcy-card-credit.luciddomains.com/index.html

These templated URLs are computer-generated based on the query and are a strong spam signal. If you look at the text on this page, you will see that it is does not make sense.

7. 100% frame

Webmasters sometimes cloak what users see by using frames. Two frames (pages) may exist, but one frame takes up 100% of the screen. The user sees one frame, but the search engine robot sees both frames.

Recognizing 100% frame:

If you are using Internet Explorer:	**If you are using Firefox:**
1. Right-click on the page. 2. Click "**Properties**". 3. Compare the URL of the frame with the URL of the page. If they are different, the page is probably 100% framed, and should be labeled as Spam.	1. Right-click on the page. 2. Click "**This Frame**". 3. Click "**View Frame Info**". 4. Compare the URL of the frame with the URL of the page. If they are different, the page is probably 100% framed, and should be labeled as Spam.

Example of 100% Frame:

http://11.freiemusik.org/khaledmp3/ This is the URL of the page.
http://www.mp3sugar.com/?aff=2607 This is the URL of the frame.

8. Sneaky redirects

A sneaky redirect takes place when a page redirects the user to a different URL on a different domain. While being redirected, you might observe the page being redirected through several URLs before ending up on the landing page. Search engines index and score the content on the first domain, yet the user is redirected to a different domain. Again, the webmaster is presenting different content to the search engine robot and the user.

One URL may sneakily redirect to a number of rotating domains, so clicking on the same result several times may land you on different pages, which may or may not look the same.

Sometimes, if you enter one of these domains into the search engine as a query, you will be taken to Amazon, eBay, or other merchants.

Recognizing when redirects are sneaky or non-sneaky

Compare the two URLs to see if it makes sense that one would redirect to the other. For example, a redirect from the old homepage for a company to its new domain is not sneaky. For example, www.compaq.com redirects to http://h18000.www1.hp.com/ in a legitimate manner. Also, redirects within the same domain are not sneaky.
If you suspect a Sneaky Redirect has taken place, you should check "who is" the registrant (or owner) of the two domains. If the registrant is the same, the redirect is not sneaky.

1. Go to the site of a "whois" provider to find out "who is" the domain registrant. Here are two you can use: http://www.domaintools.com/ or http://whois.mtgsy.net/default.php.
2. Enter the URL of each domain in the search box. (Sometimes, you will need to delete some leading or following characters. For example, if the URL is http://supportapj.dell.com/support/, you will enter just "dell.com" in the search box.
3. Compare the domain registrants for each URL. If you find that the two URLs have the same domain registrant, you will conclude that the page is not Spam. If they are different, it is probably Spam.

Example of a Sneaky Redirect:
http://www.kqzyfj.com/go65biroiq57A8E7A6577BDAA6
redirects to
http://www.jcwhitney.com/autoparts/StoreCatalogDisplay/c-10101/s-10101/TID-101?AID=1157440&PID=1428140

Using a whois provider, you will see that the domain registrant for the first URL is Commission Junction, while the domain registrant for the second URL is J.C. Whitney & Company.

Please note that "whois" may also be used when trying to determine if a page is a thin affiliate.

Spam and the Resolving Stage

It is not uncommon for Tasks to go into the “resolving” stage because Raters disagree on whether a page should be given a non-rating of **Didn’t Load**, or whether it should receive a rating and a Spam label. The disagreement occurs because Raters are presented with different pages when they click on the link in the Task. These differences may be due to Timing differences, or due to Browser and Browser Setting differences.

When this happens and the page you see matches the criteria for **Didn’t Load**, please take another look at it. Since other Raters see a Spam page, it is obvious that they are looking at something different from what you see. Here are some things you can try to change the page that you see:

1. Open the page in a different browser. If you are working in Internet Explorer, try opening the page in Firefox, Safari, Opera, and vice versa.
2. Look at the source code.
3. Look at both the live and cached pages with JavaScript disabled.

Naturally, if you do not detect spam, do not label the page as Spam. Please be aware that spam pages frequently become **Didn’t Load** pages after a period of time. If you detect Spam one day, but the page does not load for you the next day, please do not change your rating, (i.e. do not remove the Spam label).

Spam Examples on the Rater Hub

Although we have provided some spam examples in these guidelines, you can check the Rater Hub for many more examples in many different languages. The examples on the Rater Hub will also be fresher since they are checked on a weekly basis. It is also where you will be alerted to new types of spam pages that we encounter. Please make it a habit to visit the Rater Hub on a regular basis.

Conclusion

Spam recognition is a skill that is built through practice and exposure. Open discussion of difficult cases in the resolving stage in EWOQ will help you develop your skills.

When you feel unsure if a page is spam, ask yourself the following question: If I remove the copied content, scraped news feeds, fake forums and blogs, thin affiliate links, parked/expired domain links, and all that is left are PPC ads and sponsored links, the page is probably spam.

If you do not feel competent or assured enough to label a result as spam, do not apply a Spam label. We prefer that a “guilty” page remain unlabeled than an “innocent” page be labeled.

Part 5: Quick Guide to Quality Rating

The Role of the Quality Rater

As a Rater, you will evaluate '**query-page**' tasks. **Query** refers to the word or words that a user types in the search box of a search engine. The **URL** is the web address of the page you will evaluate. The **Page** or **Landing Page** is the page you will evaluate. It is the page you see after you click on the URL.

You will be given a **Task language** and a **Task location** for each Task. You must evaluate each Task in the context of its language and location. Each query will be shown in square brackets, followed by the Task language and Task location. For example, [coca cola], Spanish (MX). In this Quick Guide, please assume the query is an English (US) query unless otherwise noted.

You should understand the query before you evaluate it. If the meaning is unclear, you will need to do web research to learn about it. You will then evaluate the page based on its relevance to the query and utility to the user, and assign a rating from the Rating Scale.

Issues to Consider

Task Language and Task Location Users in different parts of the world have different expectations for the same query terms. English (US) and English (UK) users will have different interpretations for the query [football].

Multiple Interpretations Does the query have more than one interpretation? Is one interpretation the most likely or dominant interpretation? The dominant interpretation for [windows] is the universally known computer operating system. Another interpretation is a piece of glass that can be looked through.

Broad or Specific Is the query broad or specific? Broad queries are best matched by broad pages; specific queries by specific pages. [digital cameras] is a broad query. [canon SD550] is a specific query. A good result for [digital cameras] is a page with reviews about a number of different cameras. A good result for [canon SD550] is a page with a review about the Canon SD550 digital camera.

Timeliness Can a query be interpreted differently at different points in time? In 1994, the user who typed [President Bush] was looking for information on President George H.W. Bush. In 2006, his son George W. Bush is the more likely interpretation.

Query Types

A **navigational** query is intended to locate a specific web page. The user has a single web site in mind, often the official homepage or subpage of an official site. An **informational** query seeks information on a topic. The user is looking for information on the query topic. The goal is to learn something by reading or viewing content on the Web such as text, images, videos, etc.

A **transactional** query seeks to complete a transaction on the Web – for money or free – of a product or service. The goal is to download, buy, obtain, be entertained by, or interact with a resource that is available on the result page or made available through the result page.

Note: Many queries fit into more than one query type.

Rating Scale Categories

The **Vital** rating is used in special situations where the query has a dominant interpretation, the dominant interpretation is navigational, and the page to evaluate is the official web page of the query.

- Most queries do not have **Vital** results because they do not have official web pages.
- For a landing page to be **Vital**, it must be appropriate for the Task language and Task location.
- **Vital** pages may not be the best possible (i.e. the most helpful) result for the query.
- The **Vital** rating is not based on the appearance of the URL. The URL of the official homepage may contain the query terms. For example, the **Vital** result for [ibm] is http://www.ibm.com. However, www.diabetes.com cannot be **Vital** for the query [diabetes] because the query is not navigational and there is no official web page for the query, and no one can "own" the query [diabetes].
- It is possible for a query to have more than one **Vital** result. For example, [barnes and noble]: www.bn.com, www.barnesandnoble.com, and www.books.com all have the same landing page, and are all **Vital** for the query. Another example: a celebrity might have an official homepage, as well as a personal blog, a myspace page, etc. All of these pages would be **Vital**. And another example: a company acquires another company, but continues to maintain both websites. Both homepages would be **Vital**.
- Some large international corporations have country, as well as regional or global homepages. In general, the country specific homepage is the **Vital** result. If no country specific homepage exists, a regional or global homepage may be **Vital**.
- Landing pages that ask you to choose a language, country, postal code, etc., are **Vital** if the pages behind them are **Vital**.

A rating of **Useful** is assigned to pages that are comprehensive, highly satisfying, high in quality, and authoritative. **Useful** pages answer the query just right; they are neither too broad nor too specific. For many queries, they are "as good as it gets".

- If an ambiguous query has several equally strong interpretations and each possesses a unique homepage, each of the homepages would be **Useful**, e.g. [ADA]: American Dental Association; American Diabetes Association, etc.
- Other examples of **Useful** pages: an important subpage on the correct site; a page that is highly

Source: http://vizualbod.com/f/spam-guidelines.htm - 12/03/2008

informative; a timely and informative article; a page that allows the user to complete the transaction.
- If a query "asks" for a list, a directory (a collection of links) can be **Useful**, e.g. [fudge recipes].

A rating of **Relevant** is assigned to pages that have fewer valuable attributes. Relevant pages might be less comprehensive, come from a less authoritative source, or cover only one important aspect of the query.
- Examples of **Relevant** pages: a brief article on the topic of the query or a less important subpage on the correct site.
- If a query "asks" for a list, a single item is **Relevant**.

A rating of **Not Relevant** is assigned to pages that are generally not helpful, but are still connected with the query topic. The page may be too marginal in scope, outdated, too narrowly regional, too specific, too broad, etc. to receive a higher rating. They might have less information or come from a less authoritative source.

- The page may have a link to good results on the same site, but not be a good result itself, e.g. a useless subpage on the correct site. Another example is a page that has a link to good results on another site without providing any utility itself, other than the link to the "good" results on the other site.

An **Off-Topic** result has zero relevance to the query.
- The page or the URL may contain the query terms, but be conceptually off topic. For example, a page on doghouses is **Off-Topic** for [hot dog].
- If navigation to helpful content is very difficult, **Off-Topic** is an appropriate rating.

Important notes:
- A page is rated on its match to the concept of the query (i.e. how relevant or useful the content on this page is to the query). The query terms do not have to appear on the page. For example, for [paris Hilton picture], a photo of Paris Hilton is **Relevant** even if the query terms are not on the page.
- Please remember to rate the page and not the URL. You must visit the landing page and rate the content.
- If you are unsure between two ratings, go with the lower rating.

Non-Rating Categories

Didn't Load (DL) – A non-rating of **DL** applies to many different situations. Common cases are when there is no visible content, there is not enough content for the page to be evaluated, or the content is not formatted properly. Examples include: 404 error pages, completely blank pages, "page not found" pages, pages with XML or HTML only, etc. This rating is only assigned if both the live and cached pages cannot be evaluated.

Foreign Language (FL) – A rating of **FL** is assigned to a result that loads fine, but is fully in a foreign language (a language other than English or your Task language). A landing page should be rated as **FL** even if you are fluent in the particular third language. English is never considered to be a foreign language.

Exceptions:
- The page contains an image that can be evaluated in spite of the language on the page.
- The page has a link for download of software, but is understandable in spite of the language on the page.
- The page has a mix of languages, but there is enough text in English or the Task language to fully comprehend the content.
- The page has a "page not found" message in a foreign language. The rating should be **DL**.
- The page is in a foreign language, but uses deceptive techniques. The rating should be **FL** with a **Spam** label.

Unratable – Assign a rating of **Unratable** when, even after researching the query, you do not feel confident evaluating the page.

Spam Labels

Spam – When a webmaster uses deceitful techniques and web design, you should assign a **Spam** label. Please refer to Part 4, "Webspam Guidelines".

Maybe Spam - When a page appears "spammy" but you are not sure it is **Spam**, you should assign this label.

Not Spam – Assign to other pages.

Flags

Pornography – If the page has pornographic links, text, images, pop-ups, and ads, it should be flagged as **Porn**.

Malicious – If the site forces you to quit the browser, has prompts that keep coming back, attempts to download spyware, etc., you should assign this flag. Pop-ups that do *not* come back are *not* malicious.

Please note that flags cannot be assigned with a rating of Didn't Load.

Instructions

Step 1 – Research and understand the query.
- Determine the most likely interpretation(s) for the query. Is it Navigational, Informational, or Transactional? Does it have multiple interpretations? Is it broad or narrow? Broad queries are best matched with broad results; narrow queries with narrow results.
- Determine the amount of information that exists on the Web for the query. For example, a short article for a query with little information on the Web would get a higher rating than a short article for a query with a lot of information available.

Step 2 – Evaluate the page based on its relevance to the query and utility to the user. You will assign a rating from one of the non-rating categories when the page cannot be evaluated:
- If neither the live nor cached page loads, assign a rating of non-rating of **DL**.
- If the text on the page is not in English or in the project language, assign a non-rating of **FL**.

Step 3 – Assign a rating from the Rating Scale.
- Using the Rating Scale, evaluate the page according to its relevance to the query and utility to the user. Assign a rating, the appropriate Spam label and any applicable flags.
- Assign a rating of **Unratable** when, even after researching the query, you do not feel confident evaluating the page.

Part 6: Quick Guide to Webspam Recognition

The Role of the Quality Rater

In addition to evaluating a page according to its relevance to the query and utility to the user, you will assign a **Spam** label to the rating when you observe that the page has been designed using one of the Spam techniques described in this guide, or you will assign a **Maybe Spam** label to a page that appears "spammy", but that you are not confident assigning a **Spam** label.

What is Webspam?

Webspam is the term for web pages that are designed by webmasters to trick search engine robots and direct traffic to their websites. .

General Information

- Spam evaluation is done on a page basis. One page may be assigned a Spam label even if other pages on the site would not be assigned one.
- You can assign a Spam label even if the page is relevant to the query. In fact, a page can receive a rating of Vital and still receive a Spam label.
- You will assign a Spam label if you detect violations of these guidelines in any browser.
- Pay-Per-Click (PPC) ads and links appear on many pages on the web. Spammers make money when the ads or links on the page are clicked. **Please note**: Pages with PPC ads are only considered to be Spam in the absence of original content on the page. Many "good" (non-Spam) pages contain PPC ads. For example, http://www.nytimes.com/, the online version of The New York Times, a highly reputable newspaper, has PPC ads.

Spam Categories

This section describes some of the types of Spam that you will see. It is not uncommon to see more than one type of Spam on the same page.

PPC Pages

Many web pages are set up for the purpose of collecting pay-per-click (PPC) revenue without providing any or much content of their own. These pages will frequently look like search results, or they may look like a blog or message board (forum) pages.

Pages with PPC Ads only: Some pages contain nothing but PPC ads (or sponsored links).

Fake Directory with PPC Ads: With a fake directory, you will see a list of links that look like search results. However, clicking on a few links reveals that they are just ads disguised as "results".

Fake Blogs/Fake Message Boards with PPC Ads: Fake blogs and fake message boards are set up with the purpose of earning revenue from PPC ads. You will not see responses to messages, and the content is often makes no sense.

Scraped Content with PPC Ads: Scraped content refers to content that has been stolen from another source, either through the use of a piece of software that searches for content containing specific keywords, or through simple copy-and-paste. Some of the sources that are routinely scraped by spammers are: Wikipedia, the Open Directory Project, Usenet, RSS and XML feeds, and content providers who generate "articles" or other text strictly for webmasters to use to draw traffic to their pages. To determine if content has been scraped, copy a short segment of text from the page, paste it in your search box, and search for it in your browser.

Please note:

- The acquisition of content from these sources is not necessarily illegal, nor plagiarism. Webmasters who create copies usually do not claim to be original content creators and may, in fact, assign credit to the originator of the content. However, we do consider these pages to be Spam when used with PPC ads and when there is no original content provided on the page.
- Pages with song lyrics, poems, recipes, and quotations are generally not considered to be Spam, even when there are PPC ads on the page.

Parked (Expired) Domain Pages: Sometimes an expiring or expired domain is purchased by a spammer. Frequently, the content on a parked domain resembles a search directory. In addition to the directory links, the page may contain PPC ads.

Thin Affiliates: Spammers make money when a transaction is completed after the user has clicked through to the "real" merchant's site from the affiliate page.

After you click through, you will see that you are on a different URL, or right-clicking on an image on the original page may reveal the URL of the affiliate. You may also be redirected through a third-party domain. A common type of thin affiliate Spam is "hotel" spam, where you land on one page but are taken to a different domain when you attempt to complete the transaction. Amazon and eBay are also highly associated with this type of Spam.

Not all affiliates are thin (Spam). How to tell if an affiliate is thin or not:

- If there is original content added to the page, such as a review or a recipe, or if value-added services exist, such as price comparisons or coupon codes, the site is an affiliate, but not a thin affiliate.
- Yahoo! Stores are true stores, not thin affiliates.
- Small merchants who use Internet cart systems to process transactions are not thin affiliates.
- Indicators that a site is probably not a thin affiliate are shopping carts and wish lists that work, and a return policy with a physical address.

Source: http://vizualbod.com/f/spam-guidelines.htm - 12/03/2008

Hidden Text/Hidden Links: Spammers sometimes add text or links to a web page, but hide them from the human eye by making them completely invisible, using a color so that the text blends in, using very tiny text that is difficult to see, of by placing them on a portion of the page well outside the normal viewing area, for example, below the "fold" on the page.

To uncover hidden text, apply Ctrl-A to the page. This technique will display the hidden text.

Another way to reveal hidden text is by looking at the source code of the page. In IE, go to "View" and then click on "Source". In Firefox, go to "View" and then click on "Page Source". Compare the source code to what you see on page.

Please note: Hidden text is not considered to be Spam if there is no intention to trick the search engine. For example, if the webmaster "hides" the date of the update, that would not be considered Spam.

Keyword Stuffing: Keyword stuffing is the excessive use of keywords on a page. The words may be relevant or irrelevant to the query. Sometimes the keywords will be misspelled or the order of the words will be reversed. Keyword stuffing is used to draw the search engine robot to the page.

Hidden text and keyword stuffing often go together. The hidden text frequently contains keyword stuffing.

Keyword stuffing in the URL: URLs may also contain keyword stuffing. They are usually created by some type of template and are stuffed with terms that come from the query. They are often formatted with many hyphens (dashes) in them.

100% Frame Pages: Spammers sometimes frame 100% of the content from another URL.

To check in IE, right-click on the page and then click "Properties". To check in Firefox, right-click on the page, then click "This Frame" and then click "View Frame Info".

Compare the URL of the page with the URL of the frame. If they are different, you will usually assign a Spam label.

Sneaky Redirects: Spammers sometimes redirect users to a different domain. When this happens, you will notice that the URL that you clicked in the Task is different from the URL that you land on. If you watch the address bar carefully, you will sometimes observe other URLs being passed through along the way. Sometimes one URL will sneakily redirect to a number of different domains, so, if you click on the URL several times, you may end up on a different page each time.

Please note: Not all redirects are sneaky. Redirects to a different page within the same domain are not sneaky.

Also, a site might legitimately redirect from one URL to another. For example, since Compaq and Hewlett-Packard merged, the Compaq URL automatically redirects to the HP site.

Checking "Who Is" the Domain Owner

When you suspect a Sneaky Redirect has taken place, or that a URL is a Thin Affiliate for another URL, it is a good idea to check "who is" the owner of the two domains. If the owner is the same, the redirect is not sneaky. You will do this by going to a "whois" provider to find out "who is" the domain registrant. Here are several you can use:

http://www.domaintools.com/

http://whois.mtgsy.net/default.php.

You will type in the domain names and look at the information provided for each. If you find that the two URLs have the same domain registrant, you will conclude that the page is not spam.

Disabling

Spammers commonly use features of JavaScript to hide text or to show one page to the search engine and a different page to the user. You can disable JavaScript to reveal this deception by following these steps:

If you are using Internet Explorer:

1. Go to "Tools".
2. Click on "Internet Options".
3. Click the "Security" tab.
4. Click on "Custom level".
5. Scroll down to the "Scripting" section. To disable JavaScript, make sure "Disable" is selected under "Active scripting".
6. Click "OK".

If you are using Firefox:

1. Go to "Tools".
2. Click on "Options".
3. Click on "Content" or "Web Features".
4. To disable JavaScript, make sure the "Enable" box is not checked.
5. Click "OK".

After disabling JavaScript, view both the live and cached versions of the page, because sometimes the hidden text will only be revealed on one page or the other.

Spammers may also use JavaScript to create two versions of their content: one to be viewed and ranked by the search engine, the other to be seen by the user. You can also use the steps outlined above to view the different pages.

Final Notes on Spam

When trying to decide if a page is Spam, it is helpful to ask yourself this question: If I remove the scraped (copied) content, the ads, and the links to other pages, is there anything of value left? If the answer is no, the page is probably Spam.

Source: http://vizualbod.com/f/spam-guidelines.htm - 12/03/2008