Cognizant 20-20 Insights

Personalizing Search

By applying user context and uncovering essential information, search engines can deliver a more rewarding experience, resulting in more digital revenue for the organization.

Executive Summary

Billions of searches are performed by millions of users, every day. Although the numerous search engines that are currently available work well on basic keyword searches, they are often lacking when it comes to advanced research, especially to meet the requirements of information and media services providers. Regardless of the individual user needs, the search engine typically provides the same results to all.

This white paper highlights improvements that can be made to the relevance of search engine results, as well as how information and media services providers can personalize search returns based on each user's interests and requirements. It also highlights the different dimensions of personalization, and recommends ways to measure and collect insights against these dimensions.

Search Engine Results Relevancy

When gauging search engine effectiveness, two key questions must be answered:

- What should the result set consist of?
- What should the sort order of the returns be?

Search engines generally use a variation of the tf-idf (term frequency–inverse document frequency) algorithm for search and relevance (see Quick Take, next page). This means if a user searches for "the driving license," the engine takes two steps based on term frequency:

1. Find all documents with "the driving license" or a variation of that term.
2. Sort document returns, displaying those with the highest number of term mentions at the top, and the least number of mentions at the bottom.

Because "the" is a common or "noise" term in the phrase, and the other two words (driving and license) are more specific, driving/driver/drive and license/licensing/licensed will be given priority for relevance. This is called idf (or inverse document frequency). The search engine returns all documents with "driving license" and variations, ignoring "the." But the question remains: Is it sufficient for relevant search results?

Beyond Basic Search

Returning the result set based on the search term is enough for basic search engines but not for advanced research engines. Along with tf-idf, user context is the most important factor to consider. In the above example, the search engine will return all documents with the term "driving license;" however, top results might relate to applying for a license in California, which is

Quick Take

Understanding the tf-idf Algorithm

The following are variations of tf, idf or tf-idf that are used for relevance:

Variants of tf Weight

Weighting Scheme	tf Weight
Binary	{0,1}
Raw frequency	f(t,d)
Log normalization	log(1 + f(t,d))
Double normalization 0.5	0.5 + 0.5(f(t,d)/max(f(t,d)))
Double normalization K	K+(1-K)(f(t,d)/max(f(t,d)))

Variants of idf Weight

Weighting Scheme	idf Weight
Unary	{0,1}
Inverse frequency	log(N,n(t))
Inverse frequency smooth	log(1 + (N,n(t)))
Inverse frequency max	Log(1+ (max(t) * n(t))/n(t))
Probabilistic inverse frequency	Log((N-n(t))/n(t))

Simple tf-idf calculation

tfidf(t,d,D)=tf(t,d) * idf(t,D)

Recommended Weighting Schemes

Weighting Scheme	Document Term Weight
1	f(t,d) * log(N,n(t))
2	log(1 + f(t,d)) * log(1 + (N,n(t)))
3	log(1 + f(t,d)) * log(N,n(t))

Source: if-idf entry on Wikipedia, (http://en.wikipedia.org/wiki/Tf%E2%80%93idf)

> If the search engine records and analyzes the user's previous searches, however, then it would be able to apply additional context to what the user is seeking and thus provide more relevant results.

not useful if the user's location is outside of that state. As a result, top returned results are relevant only for a very small portion of the user base. If the search engine records and analyzes the user's previous searches, however, then it would be able to apply additional context to what the user is seeking and thus provide more relevant results.

In the above case, if user context is factored in, the search engine would provide California as the top result to a California resident, a New York top result to a New York resident and Alaska the top result to an Alaska user. User context typically consists of the following dimensions:

- Geographic location.
- Area of interest.
- Profession and economic background.
- Gender and age group.

Another example of how user context can be useful is when a user is located in a particular state and searches for an item, say a Samsung phone. When the results are returned with Samsung phone on top, not all the returns may be relevant for the following reasons:

- **Geographic location:** The user wants a localized search. While thousands of stores sell Samsung phones in the U.S., he is interested in just the four to five nearby outlets.
- **Area of interest:** Responses should be different for someone interested in news vs. one seeking technical details.
- **Profession and economic background:** Responses should also consider the user's profession and economic background. If a lawyer is conducting a search, for example, he might be seeking updates on an Apple vs. Samsung lawsuit[1] and be uninterested in Samsung stores and reviews.
- **Gender and age group:** Not only does interest level vary based on gender and age group; so does the relevance of results. A younger user might be seeking the best games on a Samsung phone, which differ based on age and gender.

While some user dimensions, such as gender, do not change over time, others, such as location, profession or area of interest, can vary and must be tracked and updated regularly. This is where some search engines excel over others.

For example, Netflix is preferred over other movie services because its search engine collects and acts on user context dimensions obtained through the user's profile, which includes data on:

- Age group.
- Ratings given to movies viewed.
- Number of times content is viewed (which provides area of interest and content popularity).
- Previous selections in a genre and browse history (which provides area of interest and ease of finding content).
- Viewing location.
- Amount of continuous-viewing time spent (which provides level of Interest).

As a result of a deep-dive analysis of these dimensions, Netflix can predict exactly which movie the user plans to watch during the weekend. This is the essence of what we call Code Halo™ thinking, in which companies make meaning from the intersection of digital code that surrounds people, processes, organizations and things. (To learn more, read our book, *Code Halos: How the Digital Lives of People, Things, and Organizations are Changing the Rules of Business*).[2]

Target and Context Search

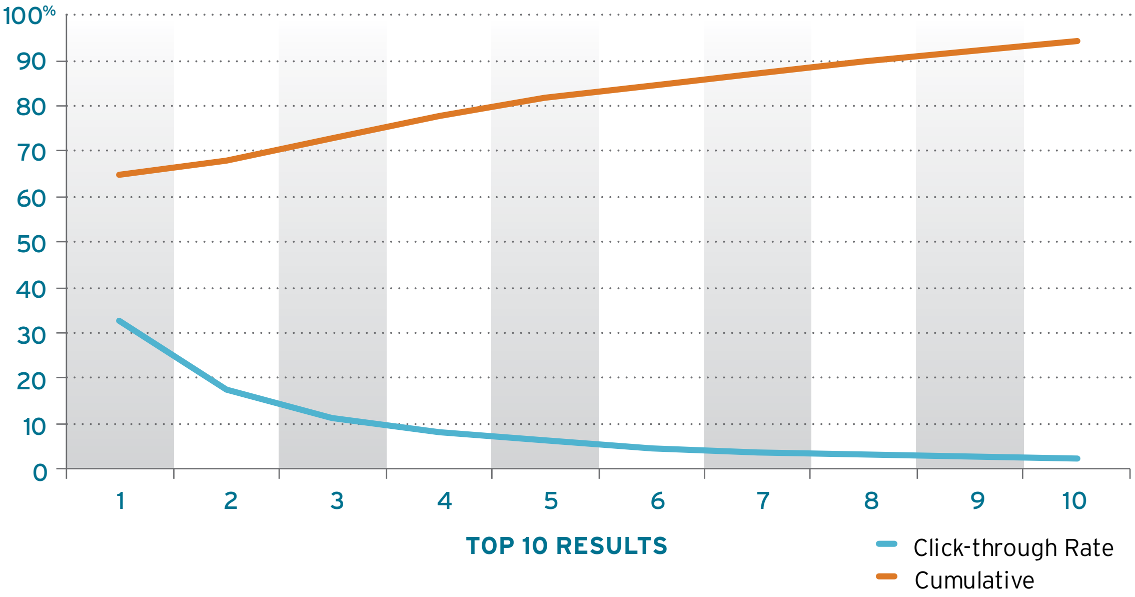
Two types of advanced search capabilities have emerged, each of which must be handled differently by the engine.

- **Target search:** Here, the user knows what he is seeking, and the engine must return specific results without any ambiguity. Even if the user incorrectly spells the search term, the search engine needs to return the right results, with supporting facts and recommendations, particularly if the user is looking for a side suggestion.
- **Context search:** In this case, the user is searching based on terms, questions or fuzzy query, and the engine needs to return the result based on tf-idf, along with user context, to return personalized results. The user should find what he is looking for within the first five results. More than 75% of click-through rates are from the top-five results (see Figure 2).

Result Ranking and Click-Through Rate

A search engine results page is the listing of results returned by a search engine in response to a keyword query. The results normally include a list of items with titles, a reference to the full version, and a short description showing where the keywords have matched content within the page.

Each page of search engine results usually contains 10 organic listings. The listings on the first page are the most important ones, because those get 91% of the click-through rates from a particular search. According to a 2013 study by Chitika, the click-through rates for the first page are:



Source: Search Engine Results Page entry, https://en.wikipedia.org/wiki/Search_engine_results_page
Figure 2

User Interaction on the Search Engine

After logging into the search engine and inputting a search term, the result set is returned. What the user does next reveals much about the engine:

- Refines/redefines the search term and searches again.
- Narrows down the result set.
- Browses the result set, clicks on it for additional details and then looks at other results.
- Browses the result set and clicks on the results that satisfy his search.
- Browses the top result set, clicks on it and obtains his required information.
- Finds his information just after submitting the search term.

Although the difference in user behavior is exceptionally small, each response actually has a very different meaning. Let's explore further.

- **Refines/redefines the search term:** When the user entered the search term, he received irrelevant results. This means the search engine did not help the user, nor has tf-idf been implemented properly, causing the user to refine/redefine the search term to find what he is looking for.
- **Narrows down the result set:** The user received results based on the term entered but also received results in which he is not interested. To clean up the result set, the user tries to remove unwanted results. Once he has the refined results, he can examine the returned documents. In this case, the tf-idf is working to its potential; however, there is no user context. Narrowing down results means the user is trying to set the right search context.
- **Clicks for more details:** The user received relevant results; however, the associated information is insufficient. To check whether each result includes the right details, the user must click and read it. In this case, tf-idf is working to its potential with some user context.
- **Browses result set:** The user received all relevant results on the first page and has the right supporting details; however, the best result is not among those listed at the top of the return, so he has to browse all the results to get to the right one. In this case, the tf-idf is working to its potential with user context. However, user context is not fully optimized.
- **Browses top result set:** In this case, the user received his results in the top result set with all the right supporting details. In this case, the tf-idf is working to its potential with fully optimized user context.
- **Finds information immediately:** In this case, the user gets his result with the submission of the search term. This is especially applicable for entity search results. This search result has all the required supporting details. In this case, the tf-idf is working to its potential with fully optimized user context.

How to Gather User Context

Every time a user logs into a search engine, he adds more data to his digital imprint, or Code Halo. This can be useful for subsequent searches, other users' searches and for the popularity of the content. When a user is new to the search engine, there is no user context.

Based on the user's interaction with the search engine, a machine learning algorithm creates user context that factors into his search pattern. When the user profile is new or inconclusive, the safest bet is to return results based on the tf-idf. Once the search engine has enough details to infer context, the result set can be personalized to address user interest. This not only ensures that more relevant results are returned, but it also saves unnecessary user effort.

Based on the user interaction with the search engine, a machine learning algorithm creates user context that factors into his search pattern.

User context is a continuous and ongoing learning process. For some users, there may not initially be a conclusive pattern, but over time, a specific pattern may develop that yields user context. In other cases, the engine needs to unlearn some patterns and create a new pattern with updated user context (see Figure 3, next page).

Context can be drawn from the user's previous searches, as well as from users with similar interaction patterns.

- **User's previous searches:** The safest way of creating user context is to refer to previous user behavior with searches and intuit a research

The User Context Learning Loop

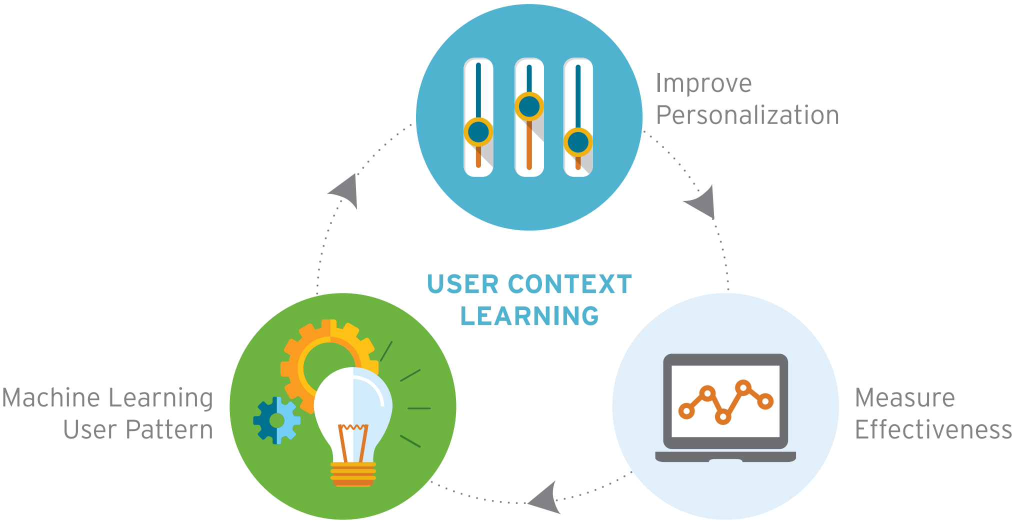


Figure 3

path. Once there is clarity on user context, the engine provides relevant results and suggestions. For example, iTunes applies Code Halo thinking to help the user discover new music based on what she already has or likes. It relies on the user's context and recommends new music.

- **Users with similar interaction patterns:** This model uncovers similar users' research patterns and applies those learnings to return more relevant results (i.e., collaborative filtering[3]). For example, an "other topics you might be interested in" feature is based on other users' similar patterns. This approach may not be used by all search engines, especially paid engines. Users may raise concerns that others benefit from their search pattern. This option should be used on a case-by-case basis; for example, while it should be avoided in legal or scientific research, it can be very useful for music streaming websites.

Even if no specific user context is obtained, context can be derived from the popularity of the content.

Looking Ahead

General search engines take the user's entered search term, remove the noise, apply tf-idf and return the result set based on the relevance order. The result set may or may not be useful to the user, and relevance is dependent on user context. To make search results more relevant, information and media organizations need to:

- Measure the effectiveness of the search engine.
- Understand the user's search pattern.
- Relate the user pattern to specific dimensions of the user context.
- Apply the user context to personalize the results for the user.

The above steps should be a living, breathing, continuously evolving process. This will help organizations return personalized results, and deliver content that improves the overall satisfaction with the search engine experience.

Note: Code Halo is a trademark of Cognizant Technology Solutions.

Footnotes

[1] Wikipedia entry on Apple vs. Samsung, https://en.wikipedia.org/wiki/Apple_Inc._v._Samsung_Electronics_Co.

[2] *Code Halos: How the Digital Lives of People, Things and Organizations are Changing the Rules of Business*, by Malcolm Frank, Paul Roehrig and Ben Pring, published by John Wiley & Sons, April 2014.

[3] Wikipedia entry on collaborative filtering, https://en.wikipedia.org/wiki/Collaborative_filtering.

About the Author

Utsav Joshi is an Associate Director within Cognizant's Information, Media & Entertainment business unit. He has more than 16 years of service delivery experience, with a key focus on global project and service delivery management and an emphasis on architecture and delivery management in North America. He is currently responsible for program and solution delivery of an online legal research application developed by a leading U.S.-based legal information services provider. He has successfully delivered large critical programs for Fortune 500 organizations across multiple segments. He holds a B.E. (electrical) and is a TOGAF and AWS certified professional. He can be reached at Utsav.Joshi@cognizant.com.

About Cognizant

Cognizant

World Headquarters
500 Frank W. Burr Blvd.
Teaneck, NJ 07666 USA
Phone: +1 201 801 0233
Fax: +1 201 801 0243
Toll Free: +1 888 937 3277
Email: inquiry@cognizant.com

European Headquarters
1 Kingdom Street
Paddington Central
London W2 6BD
Phone: +44 (0) 20 7297 7600
Fax: +44 (0) 20 7121 0102
Email: infouk@cognizant.com

India Operations Headquarters
#5/535, Old Mahabalipuram Road
Okkiyam Pettai, Thoraipakkam
Chennai, 600 096 India
Phone: +91 (0) 44 4209 6000
Fax: +91 (0) 44 4209 6060
Email: inquiryindia@cognizant.com