

B E Y O N D

Think Beyond. Think Mphasis.

WHITE PAPER



Managing the Information Overload – a Findability Solution

Kunal Parekh,
Enterprise Search COE

January, 2011

Table of Contents

Defining the problem	2
Welcome to a world of information overload	2
Managing the information overload	2
Linguistic Relevancy	2
User Intent	2
Reverse Intelligence	2
Meta Tagging, Folksonomy and Taxonomy	3
Personalization	3
Conclusion	4
Suggested Reading	4

Defining the problem

A decade ago, when people wanted to search for information, it required either contacts or intellect or both on how to get the correct relevant information. Google changed that forever. The simplest way to obtain information today is to “Google it”. They introduced a search engine with a completely new paradigm with new search algorithms, which would rank pages based on how many back links or references the page has on the web. This gave better results, but people still landed up with hundreds and thousands of results. There is too much information on the internet, most of which is either irrelevant or duplicated. Thus there is an “information overload”¹. It is a term coined by Alvin Toffler which means to have an excess amount of information to make a single decision. It is excessive information which will confuse the person instead of solving the problem. Let us take an example. The other day, while searching for “travel insurance from Mumbai to Lyon”, 332,000 results came up on Google, 502,000 results on Bing, 568,000 results on Yahoo. The expectation was to get a maximum of 5 results, related to Indian insurance companies providing travel insurance for these sectors. With all this information, it has created more confusion rather than solving the problem.

Welcome to a world of information overload

The purpose of searching for information is lost because of the way search engines work. Today, people need a new kind of intellect to get the needed information. They need to understand keywords and the order in which they should be keyed in, as both of these will affect the resulting output, which is still a huge amount of results. This brings us back to the problem statement, what information out of the result is actually relevant to them? How to get more relevant information with fewer results? How to get results based on the user’s intent or context and not based on a search engine’s algorithm? How to find solutions to all these problems?

Unluckily these problems are not limited to the web but also proliferate within the enterprises. Today, enterprises have a variety of content and data sources, applications, and collaborative places and these (may) differ across regions. In the enthusiasm to implement findability across the enterprise, the enterprise generally makes all the intranet information (based on their access authorization) along with other related website information available to the user. Now, the user is faced with the same problem as the web - getting a load of results from a search. IDC estimates the information workers spend on average 48% of their time searching for and analyzing information (9.5 and 9.6

hours per week, respectively) which costs an organization \$28,000 per worker per year². How to address these practical issues faced in an organization? Which systems to index? How much of it should be indexed? How to harmonize data across systems? How to help employees to get the information they are searching for? How to optimize the results? How to understand their context or intent? These are some of the concerns and questions raised by most of MphasiS’ clients.

Managing the information overload

Linguistic Relevancy

The content within an enterprise will grow year on year, so how to optimize the results? Linguistic relevancy is a common way to resolve this issue. Most of the search engines provide this feature out of the box with their proprietary relevancy model of ranking data and content. The search engine performs linguistic analysis (is a process of transforming content in order to improve relevancy, recall and precision.) on the content and returns a ranked result set based on this analysis. Linguistics processing helps in defining the right words that can be associated with a document thereby increasing the relevancy of the documents. While this is the regular way of resolving the information overload problem, it still continues to give us a lot of irrelevant content as the relevancy models may not understand the intent or the context of the person searching the same.

User Intent

As a next step, the content could be sliced and diced based on the user’s intent (the intent or context with which the user performs a search). Categorizing the content based on intent is a good way to improve findability and reducing the results returned. Let us take the example of Amazon where they have categorized the information based on the intent of the user. Each intent is a department like books, videos, magazines, etc. The user can easily select the intent of his search and limit the results to his / her intent. There are a lot of sites using this method very successfully.

Reverse Intelligence

If the above methodologies do not suffice, then the next level of refinement can be implemented. With the relevancy and intent in place, a certain set of ranked results based on the search relevancy and intent will come out. However, it will always return the same ranked results for the same set of keywords searched. The reason for this is that the search engine has no channel of receiving feedback of the users’ usage. This means that there is no way to understand the usage of people like the keywords being searched for or the content viewed for the searched keywords.

¹ Wikipedia: Information Overload - (http://en.wikipedia.org/wiki/Information_overload)

² The Hidden Costs of Information Work, IDC April 2006

Using reverse intelligence, this information is collected and fed back to the search engine, thus completing the cycle. In this step, it is possible to boost the content or the documents that were viewed or downloaded with the keywords they had been searched for. The results are never static as the user's usage has a role in ranking the results. The set of results clicked most often for a set of keywords will be boosted to the top and the remaining results will follow. What is done using this method is change the ranking based on user's feedback.

There are a few intricacies that need to be taken care of using this approach and an incorrect solution will generally take away the competitive advantage of this approach, making it an anti-pattern.

Case Study: One of MphasiS' clients had implemented search within their enterprise. They wished to implement two new features as top keywords searched and zero results keywords. During the requirements, it was proposed that along with these two features the client should also implement the reverse intelligence model, where the results also would be ranked based on the views/downloads the users performed on keywords searched. They agreed and were very happy with the resulting implementation. Now they had a search application which could learn as to which search results were more important for certain set of keywords.

Meta Tagging, Folksonomy and Taxonomy

The constraint with the reverse intelligence solutions is the difficulty to understand the context of different user groups searching for the same keywords. For example Sales person searching for "Interwoven" may be searching it with the context of the costing of the licenses or feature list while a Developer searching for the same word "Interwoven" may be with the intent of finding the Developer's Handbook.

To resolve this issue, the author generally adds metadata to the documents for the purpose of findability. However, it is often observed that the tags added by the author are from his / her perspective and for the purpose of document or content management. The meta-tagging would depend on one person's ability to identify tags for the document.

To overcome this problem, a new concept called folksonomy or social tagging was introduced. Tags would be either keywords or phrases or attributes which would make the content more relevant from a retrieval perspective. In this method, users can add their own tags to the content. Tagging helps both the user tagging the content and the other users searching for content as they will now be able to retrieve the same content with the additional keywords added. This makes the content more usable as it is now retrievable with more keywords.

It is argued that a user can tag the content only once he finds the document. This is certainly true but when looked

at from an enterprise perspective, it is only the first person who invested a longer time on finding the document while was not tagged. All the users after him will reach the same content in lesser time due to the additional tags. So in the long run, tagging will help in reducing the time spent on searching the right documents.

Case Study: One of the clients at MphasiS had a search application with a lot of data sources (with a total data size of approximately 3 TB). The problem the users faced was with locating certain category of documents (e.g. white papers against technical papers) in a huge set of search results. The other problem they faced was locating the same document again later in time. MphasiS advised the client to implement a taxonomy which the company related to and request all content owners to categorize their content using one of the taxonomy keywords. The other advice given was to allow the users to tag the results with the keywords they related to. Over time, the users were happier with the findability of the documents and found them relatively quickly by using these features.

Personalization

Another solution that can be implemented in reducing the information overload is via content personalization. In this method, the user has the ability to create his personalized settings. The user can set things like in which sources he would like to search (like corporate company news, intranet, knowledge base, etc.), what type of content he is interested in (like white papers, technical articles, etc.), profiling his personal data (like demographics, interests, age, sex, etc.). A lot of personalization combinations are possible which will vary from company to company. To implement these, an advanced search (with save as default) feature could be provided to the user, where the user can choose between the various personalization or another way could be to give these as filters to narrow the scope of the search (with save as default).

Conclusion

Different ways to manage information overload and how to organize the information to be more intuitive, contextual and relevant have been presented. These are the methods which MphasiS implements for most of its clients to achieve the ROI on their requirements.

For any implementation, the right mix of these methods is necessary to get the maximum ROI on the solution and each findability solution varies based on its requirements. While there is no fixed formula on how much of each method should be used, there is an optimum balance between each of the methods which varies from organization to organization.

Suggested Reading

“Surviving Information Overload: How to Find, Filter, and Focus on What’s Important”, Odette Pollar, published by Crisp Learning (2003), ISBN 978-1560526940

“Social tagging as a classification and search strategy: A smart way to label and find web resources”, Serena Bonino, published by VDM Verlag (2009), ISBN 978-3639207743

“Keeping Found Things Found: The Study and Practice of Personal Information Management”, William Jones, published by Morgan Kaufmann (2007), ISBN 978-0123708663

Contact us

USA

Mphasis
460 Park Avenue South
Suite # 1101, New York
NY 10016, U.S.A.
Tel: +1 212 686 6655
Fax: +1 212 686 2422

UK

Mphasis
88 Wood Street
London EC2V 7RS, UK
Tel: +44 208 528 1000
Fax: +44 208 528 1001

AUSTRALIA

Mphasis
410 Concord Road
Rhodes, NSW 2138, Australia
Tel: +61 290 221 146
Fax: +61 290 221 134

INDIA

Mphasis
Bagmane Technology Park
Byrasandra
C.V. Raman Nagar
Bangalore 560 093, India
Tel: +91 80 4042 6000
Fax: +91 80 2534 6760

About Mphasis

Mphasis is a global service provider with \$1B in revenues, delivering technology based solutions to clients across the world. With currently over 39,000 people, Mphasis services 494 clients in Banking and Capital Markets, Insurance, Manufacturing, Communications, Media & Entertainment, Healthcare & Life Sciences, Transportation & Logistics, Retail & Consumer Packaged Goods, Energy & Utilities, and Governments around the world. Our competency lies in our ability to offer integrated service offerings in Applications, Infrastructure Services & Business Process Outsourcing capabilities. We are uniquely positioned to service our clients with best cost-performance. To know more about Mphasis, log on to www.mphasis.com