

# Fighting Information Overload with Agents and Personalization

---

---

---

## Table of Contents

|   |    |
|---|----|
| 1. Introduction .....                       | 3  |
| 2. Push vs. Pull? .....                     | 4  |
| 3. Retrieval today, Delivery tomorrow ..... | 6  |
| 4. Too much information .....               | 7  |
| 5. Agents to the rescue .....               | 9  |
| 6. Information appraisal .....              | 10 |
| 7. Personalization .....                    | 11 |
| 8. Putting it all together .....            | 13 |
| 9. References .....                         | 14 |
| 10. Authors.....                            | 15 |
| 11. Copyright .....                         | 16 |

---

## 1. Introduction

There is little doubt that people will soon be overwhelmed with the amount of information exposed to them. We already see manifestations of this today, with the feeling that we are drowning in information but starved for knowledge slowly emerging. In 2006, approximately 161 exabytes (one exabyte equals  $10^{18}$  bytes) of information were created. That's estimated to be more than in the previous 5,000 years! Credible predictions put this amount at 988 exabytes in 2010. Some problematic implications for our future are obvious. People will need advanced tools in order to handle such huge amounts of data. How will we manage the problem of information overload?

It was realized some time ago that information growth is a problem but we are still very much in the first stages of the information age. Currently most of the effort in addressing this issue is going into "managing" the information, the focus being mostly on storage, classification and retrieval. Examples of this are context sensitive search, page ranking, cloud tags, natural language processing etc. While necessary, these are not permanent solutions to the problem we face in the long run. In the future, are we doomed to spend our days filtering "search results"?

This paper will consider the implications of the information explosion that is already upon us and will only magnify in the coming years. We will explore how we are getting our information today, consider how this is likely to change in the future and why *agents* will play a large role. Finally we will discuss *personalization* as another dimension that we believe to be a necessary component in a long-term solution to information overload.

---

## 2. Push vs. Pull?

Let us consider the two main methods we have of getting information.

The *Pull* method involves initiating retrieval of some well-defined information from a source, which we will refer to as a content provider. No long-term relationship is created with the content provider as this is a single transaction.

Pulled information usually satisfies some immediate or short-term need, pulls are usually performed in an ad-hoc fashion. It is information we only want when it is useful to us and is usually disposable after we have seen it. Furthermore, it is likely to change frequently. Examples of this are weather forecasts and currency rates. For future reference we will call this kind of information on-demand.

The following assertions can be made with regards to pull and on-demand information:

- a) Consumer already knows information exists
- b) Consumer already knows how to find information

The *Push* method is different as information is pushed from content providers to consumers without any specific information being requested. For this to be possible, a long-term relationship of some sort usually has to be created between the content provider and the consumer, we will refer to this as a subscription. In this case, long-term simply means that it is valid until either the consumer or content provider terminates it

The nature of pushed information is that it satisfies criteria that the consumer previously stated when the subscription was created. These criteria may be as simple as the identity of the content provider. Pushed content satisfies a perceived future need for information of the consumer. This information is usually not disposable, it does not become obsolete in the near future. Pushed content takes many forms, such as newsletters, magazines, TV stations, podcasts and RSS feeds. We will refer to this kind of information as *passive*. Pushed content is normally moderated by editors. In that case the subjective view of the editor influences the content selection.

The following assertions can be made with regards to the push method and passive information:

- 
- a) Consumer has found content provider
  - b) Consumer has created a subscription with content provider
  - c) Consumer does not know information exists until received
  - d) Information is selected by editors

Before we continue, we should note that pull and push are not competing methods. It is meaningless to consider which is better or worse without also considering if we are dealing with on-demand or passive information.

---

### 3. Retrieval today, Delivery tomorrow

Surveys imply that awareness of push methods is as low as 12% among Internet users and that only 4% have knowingly used them. Obviously, people cannot use that which they do not know exists. It is clear from this that pull is dominant today, regardless of what kind of information we consider. Why is that?

There are two main reasons for this. The first is that most communication protocols are pull based, especially the most popular ones on the Internet with HTTP being a prime example. There can be no direct push from the web to a user. This is not to say that these protocols must be replaced for push to be possible, rather that the web was designed as a pull-centric system. The second is that there hasn't been a great need for popular push protocols in recent years because we have been able to find and retrieve the information we need ourselves.

Today, search engines help us find a lot of the information we need and we remember the bookmark or remember the URL's of our favorite websites. We assume that by regularly monitoring a manageable number of web sites we won't miss new information relevant to us. But even today, how realistic is this assumption? This certainly won't be feasible in the future with information growing exponentially all around us.

Editorial selection of material is also an issue worth considering. The web sites we monitor today are our content providers. They feed us information they believe is relevant, but the very fact that someone is selecting what information we see increases the chance that we will miss information we need. The only practical way to manage this risk is monitoring more content providers, but this would take even more of our already limited time.

The larger point is that pull methods can only continue to dominate as long as we are finding it feasible to find the information we need ourselves. If there comes a time when this is not the case, the popularity of push will increase dramatically.

In the next chapter, we will discuss why this situation is "around the corner".

---

## 4. Too much information

As we touched on earlier, the rate at which new information is being produced has been growing exponentially in recent years. It is hard to envision that this rate will slow down in the next 10 years and even harder to envision that it will ever drop below what it is today.

The fact that this vast amount of information is being produced is not a negative thing, rather an indication of progress and invention. However, it has troubling side-effects. As available information increases, it becomes harder to find the information we need.

Search engines are currently our preferred method of finding information. They have been making improvements in recent years and will no doubt continue to do so in the future. They are a significant reason why pull is dominating over push today. But will they be up to the task of guiding us through the digital jungle of the future?

Let us proceed by considering the following questions.

Is it unreasonable to assume that as the Internet grows in close relation to the growth of information that:

- a) Search engines will have trouble keeping up and doing traversals of the Internet in a reasonable timeframe?
- b) Our search queries will have to become ever more elaborate and complex to be as effective as they are today?

We find that these assumptions are not unreasonable considering current trends. It is clear that search engines face considerable challenges in the coming years. In any case, they will almost certainly become more troublesome to use as information increases. They will continue to be the main source for pulled content, but what constitutes pulled content will change.

The growth of information has already been discussed. This is an intuitive, logical and perhaps obvious trend that is supported by scientific predictions. There is nothing, on the other hand, to credibly support that the mental capabilities of the average human will increase in the immediate future. Our "bandwidth" is fixed, but what's coming in "over the wire" is growing exponentially.

---

Thus, the feasibility of pull methods is likely to decrease dramatically in coming years and push methods are likely to become far more popular or even dominant. As we are dealing with exponential changes, we must also keep in mind that this will happen quickly. With the haystack exploding in size we must quickly implement ways of finding the perpetual needle or risk losing it forever. One may argue that search engines will also improve as this happens, but no one can argue that with increased volume of produced information we risk missing more and more information important to us.

---

## 5. Agents to the rescue

What is needed is a solution that finds the information we need for us. We won't be productive if we have to find it ourselves, it will take massive amounts of time and patience. Only with this kind of solution will we truly overcome information overload.

Let us consider what qualities a long-term solution must possess:

- a) Individual oriented, different people need different information
- b) Multi-source, consumes content from many content providers to overcome editorial bias
- c) Must constantly monitor new content from content providers for any new information the user needs
- d) Must be able to adapt, as user preferences will change over time

Considering the nature of this proposed solution, the term *Agent* seems appropriate. We are envisioning an entity that is constantly looking for new information you need and lets you know when it is found. Meanwhile, people will be free to concentrate on more productive and hopefully enjoyable things. Agents will find the needles in the growing haystack. By constantly adapting to your changing needs, they will know what kind of information you need and their sole purpose will be to find it.

It is clear the challenging part of creating such a solution is capturing individual preferences and applying them to information. We will explore this issue in a later chapter. Less clear is how to quantify to what degree a given piece of information "fits" the preferences of a certain user.

---

## 6. Information appraisal

Let us consider the task of fitting information to individual preferences from a slightly different perspective. We are in fact trying to determine if a user will need or be interested in some piece of information he has not yet seen. Such information can be said to be of value to the user, but bear in mind that any two individuals will almost certainly value the same piece of information differently. Quite logically, we will call this *Information Value*.

Having established this, let us consider our task again. It now becomes clear that we are in the business of *Information Appraisal*. The market is booming but it is very much a buyer's market. Our task is equivalent to answering how much a piece of information is worth to a specific individual. In other words, we are attaching a personalized value to the information.

Information that is very useful to us will have high value. Less useful or interesting information will have lower value. Information that is not at all useful and perhaps irrelevant to us will have zero or negative value because consuming it wastes our time, and to quote a tired cliché: time is money.

In this information market, agents will be our brokers. They will aim for the highest commission but not collect it.

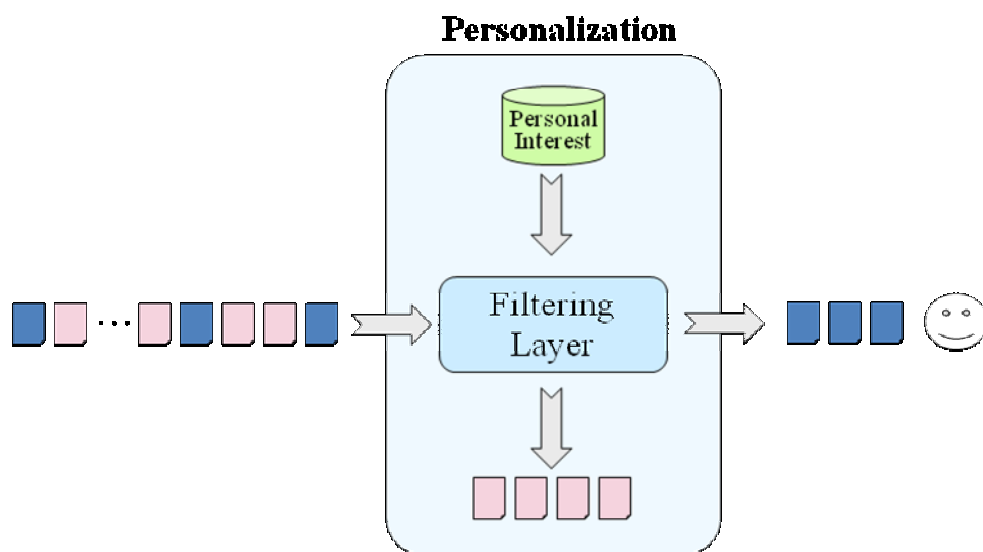
To clarify, we are not referring to any actual money changing hands; the concept of value is just most tangible as monetary amounts and was chosen for this reason. We believe the general ideas described in this chapter and the concept of information value to be most useful in guiding our efforts towards fitting information to individual preferences.

Statistics show productivity in the US business sector growing by 2-3% every year for the last five years. For this trend to continue, it will become necessary to minimize time that is "wasted" on searching and filtering information while maintaining or improving the current accessibility to the information we need.

---

## 7. Personalization

To create a personal relevancy layer, the largest task is capturing the preferences of an individual. Clearly, preferences are elusive things that are hard to capture. They exist in our minds and change over time. Further complicating things, current technology limits us more or less to a mouse and a keyboard. Is it possible to convey something so complex through such a rudimentary interface?



*Figure 1 – Personalization in action*

Features are a key concept in this regard. Information at its most basic level of zeroes and ones has features, such as reoccurring patterns or a lack thereof. When a specific type of information is considered in this regard, more meaningful and intuitive features emerge. In written text it might be the occurrence of certain words. In audio and video it will be something very different.

With the current state of technology, we believe the optimal solution to this task is a two layered approach:

### 1. Preference dictation

If the features are commonly well known and understood the user can *dictate* them. This requires the user to be able to articulate, understand and be aware of his preferences. Preference dictation might not be suitable for all types of information. It is not easy to envision how this would work in practice for

---

multimedia information for example. However, this is easier with written text because it has easily understandable features such as words, sentences etc.

This approach is very direct and intuitive. But bearing in mind the rudimentary nature of the interface, this is only practical for a small portion of anyone's preferences. And as we have mentioned, for certain types of information, it might not be practical or possible at all.

## 2. Learning

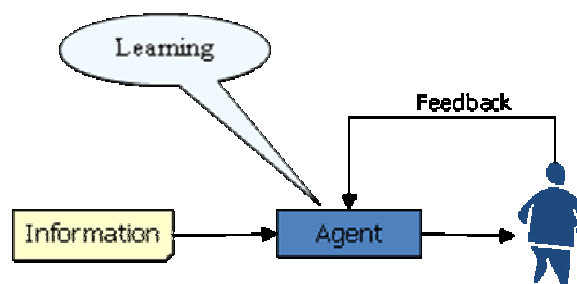
Given a mechanism to indicate whether some information is useful or irrelevant, it is possible to learn from an individual to gradually shape his *Preference Landscape*. There needs to be no understanding on his side of what the features are or what they represent, he is simply indicating something about the information as a whole (and all of its features).

By preference landscape we refer to some kind of framework or mechanism for storing preferences. How this is implemented is beyond the scope of this document.

Over time, constructed preferences can tell us what features the individual likes and dislikes with an acceptable level of accuracy.

The power of the learning approach is that it is able to capture very complex preferences and even subconscious preferences that can influence individual's favourism or decisions. Over time, constructed preferences can tell us what features the individual likes and dislikes with an acceptable level of accuracy.

With preferences changing over time, it is clear these methods will have to be continuously applied using user feedback.



*Figure 2 – Agent learning*

---

## 8. Putting it all together

Information is growing, fast. We are currently finding the information we need ourselves. As information increases rapidly, this will not be a feasible and long term solution to our thirst for knowledge. If we produced more information in 2006 than in the previous 5000 years, and are expected to produce about 6 times more information by 2010, it is obvious that changes will be necessary. We will be forced into a situation where we will rely on information to be delivered instead of finding it ourselves, this will shift our focus from pulled to pushed information. In fact, we are already seeing this happening with the growing popularity of RSS feeds. However, we will still not find it manageable to find all the information we need as the information providers will also send more and more information our way.

Connecting individuals with the information they want will become a critical task that must be automated. This is where agents come into play, software entities that know your preferences and can decide if information is relevant to you. They will provide us with an affordable way to stay productive while still seeing the information we need to see.

These agents will have to know your preferences and with current technology they can do so by dictation and learning. More sophisticated technology will make this task easier in the future, but it can already be accomplished today. Using personalization methods inspired by information appraisal, agents will know what is most relevant to you.

As these changes are happening at exponential rates, our timescale to react to changes is compressing. We have no choice but to keep up and ever decreasing time to adapt. With the future just around the corner, information overload is creeping in. We feel better to have an understanding of what is necessary to overcome it.

---

## 9. References

The Expanding Digital Universe – A Forecast of Worldwide Information Growth Through 2010

[http://www.emc.com/about/destination/digital\\_universe/pdf/Expanding\\_Digital\\_Universe\\_Executive\\_Summary\\_022507.pdf](http://www.emc.com/about/destination/digital_universe/pdf/Expanding_Digital_Universe_Executive_Summary_022507.pdf)

RSS – Crossing into the mainstream

[http://publisher.yahoo.com/rss/RSS\\_whitePaper1004.pdf](http://publisher.yahoo.com/rss/RSS_whitePaper1004.pdf)

Private business sector: Productivity & related measures, 1987-2005

<http://www.bls.gov/news.release/prod3.t01.htm>

---

## 10. Authors

Pejman Makhfi is a Silicon Valley technology veteran, serial entrepreneur and angel investor in the high-tech industry. Pejman has more than fifteen years of progressive experience in providing consultancy services and best practices to entrepreneurs, technology investors, and forward-thinking Startups.

Helgi Páll Helgason is a software engineer and artificial intelligence expert holding B.Sc. and M.Sc. degrees in computer science from the University of Iceland. Helgi has over 9 years of experience working in various fields such as investment banking, anti-virus and bio-tech.

---

## 11. Copyright

This paper is copyrighted by and distributed through [Perseptio.com](https://Perseptio.com), but it may, by the permission of the authors, be freely downloaded, translated, printed, copied, quoted, distributed in any appropriate media providing only that it not be altered in any way in text or intent and that the authors are properly credited.