

Measuring the Efficiency of Documents Retrieval

Abstract

The design of the management of document storage and retrieval is key for a successful business. Documents need to be saved to a safe location and retrieval must be efficient. This article compares approaches to searching for finding documents stored in the business and makes an estimate of the time with grand search and tag-based designs.

Introduction

How do users of computers find files? Those of us blessed with a good memory and a tidy approach to work will store files with sensible names and in a sensible location and then when the need retrieve the document arises do so without too much fuss at all.

However when saving a file it is not always obvious what the correct name or location for that file should be. A document may relate to more than one customer and to a project as well. Back a couple of decades the secretary would file a letter in one Filecabinet and file cross-references in the other potential locations. A filing approach based on folders and sub-folders, say for each customer, starts out appearing very logical and simple to use, but quickly falls into disarray. Documents are often duplicated as different users decide to classify them differently.

So what do we do when we can't find a document? Many of us will not even remember where we have stored a file or indeed what the file is called after the passage of a few weeks or indeed months. Add to this the difficulty of guessing what location our colleagues used and the size of the problem is clear. Finally there may be no bigger cause of frustration in the office than staff saving documents on their desktop or in **My Documents**.

Search or Tag?

There are a number of approaches that can be used to find lost files. Windows provides its own search facilities allowing you to find documents both by title and even content. This general search capability may be enhanced by other general search facilities including for example Google's Search Appliance offering.

Another approach may be to be more disciplined when saving the file in the first place. This approach allows users at the time the document is saved, and when they have the best understanding of its content, to save metadata or tags along with the document describing the content. Tags may include for example the customers or suppliers that the document relates to, the project and any number of keywords. A key feature is that a document can be tagged with many customers effectively placing the document in more than one folder.

The grand search approach, perhaps best implemented through the Google Search Appliance, has the great feature that it will work despite the user. There is no requirement on the individual saving the file to categorise the document in any way. However the user trying to find the file is left attempting to find search terms that best describe the document in an adequately unique way. As we all know this is difficult.

The more disciplined tagging approach does require the user saving the document to consider the best tags to describe the file and certainly requires some diligence on the part of the author. However it is the author that can best categorise and tag the document. If time is invested up front to provide these tags then finding the document later is likely to be simpler and quicker. In some sense the decision needs to be made whether to invest time as documents go into the repository or later as users try to find documents.

Allied to the tagging process can be the management of the naming and saving to the correct location. As a user saves a file they enter the tags and then the document is taken off and saved to the appropriate location on the file server.

Measuring the Efficiency of Documents Retrieval

Measuring the cost of searching

Firstly we should state that the cost of never finding a document may very large indeed, but perhaps hard to calculate. What follows is a calculation of the cost of finding a document based on the general search and searching against tagged files.

If you analyse the time spent by the author and subsequent readers of the document it is possible to try and measure the approach that is most efficient:

Time to conduct a search and look for the document of interest within the found list	10 seconds
Time for an author to tag a document	5 Seconds
Success rate for grand search	20%
Success rate for tag-based search	75%
Documents written per day	3
Documents read per day	20
Time used in search by 10 members of staff	
Grand search approach	2.8 hours per day
Tag-based search approach	0.8 hours per day

For an organisation of 10 staff the tag-based approach saves perhaps 2 hours a day – 0.8 hours as compared with 2.8 hours. For 100 staff this climbs to 20 hours or between 2 and 3 staff member's time.

In conclusion

It is clear that a disciplined approach to storing and subsequently finding documents can be one of the most important procedures designed for a business. A complete failure to address this issue can lead to enormous cost and frustration. The primary task is to ensure the file is saved to the appropriate location on the server where it is accessible to all and will be backed up.

Tagging the documents as they are saved offers significant advantages over relying on a general search to subsequently find documents. We have attempted to quantify the impact on the organization of these two approaches and have shown considerable savings in adopting a tag-based approach.

Ian Manning, Baycastle Software Ltd.