Legal Information Management, 10 (2010), pp. 24–28

doi:10.1017/S1472669610000071

Search Engine Optimisation and Automatic Classification

Abstract: Derek Sturdy explains the importance of search engine optimisation for the legal information professional involved in the organisation's website in the Google era and suggests that the most important pieces of information are the title and the abstract. He also discusses the rise in automatic classification in the enterprise search context.
Keywords: search engine optimisation; automatic classification; enterprise search; websites; automatic indexing

Introduction

In 2009, I had to tackle two totally different jobs, which produced similar conclusions. Until then, I had not realised how connected the questions of automatic classification and Search Engine Optimisation have become. I thought that they had little to do with each other. I thought that SEO was something done by people selling, say, gadgets or holidays, to try and get their website further up in Google search results lists than the endless e-Bay advertisements; worthy, important, but not much to do with knowledge management in law firms, for example. I thought that automatic classification was a sop, a sort of bone thrown patronisingly, by software giants like Autonomy, to the poodles of Information Services in client firms, to make them feel better, while actually the mighty search engine just did it its way, regardless. I found out that the reality was quite different. This contribution tries to set out what I have learnt.

Search Engine Optimisation

A side note on the commercial aspects

SEO means that you want to make your site come up in the first five results on Google. Despite the protestations of the innumerable services and companies which promote their uniquely clever ways to achieve SEO for you, these outfits always seem to be one jump behind Google, (and the other engines). Why? Because the search engines change the algorithms and rules all the time; but they never tell you just what the rules are.

For people who are serious about selling their goods and services on the web, there is, therefore, one fundamental and enduring rule in commercial web SEO: it only works well if you pay. If you want to be in the top five on Google, you have two choices: offer something so esoteric that there is no competition, or get out your cheque-book. The bigger your market, the more you will have to pay. Do not pay, and after a few days when you think "Aha, cracked it!" you will suddenly discover that the first reference to your material is #525 in the list – i.e. invisible.

You probably knew that already. But valuable, and expensive, as I found that lesson, that is not the point for us.

The middle ground: attracting clients to your organisation's website

What about the various articles and pieces that you post on extranets and websites, to show clients that your organisation has the legal expertise, so they should keep using you, or move to you? Promotional know-how and regulatory updates on websites are among the most effective advertising used, for example, by law-firms in emerging jurisdictions such as South East Europe or Central Asia, but this is an important issue for everyone in legal information. When a client searches for online advice, you want the appropriate content from your organisation to be up there in the top five results. You are like the people who need commercial SEO, but (almost certainly) without the budget. The critical difference in your audience is that, in contrast to internet shoppers, most of your intended audience already use, or will soon use, enterprise search applications, large or small. This is important.

You can help by getting three things right. I will offer some explanations, and then some practical tips. They are (in this order):

- The title
- The abstract or first paragraph
- Hypertext references

Once you take the "paid for" aspect out of it, external search engines rely heavily on the title. Why shouldn't they? The title is supposed to tell the punters what the text is about, so it seems reasonable to use it. It is normally much the most important part of a results list. It is also the one piece of metadata that is completely open. You cannot get more hits by lacing your title with pornographic words, or endlessly repeated catchphrases in every conceivable combination – the title is right up there, in front of everyone. What your users hope is that the content of your work is accurately described by the title.

The next most important piece is the first paragraph, or the abstract. This is particularly important where your clients use some form of enterprise search (as most substantial companies do). These applications pass search queries to the various external engines, but what they get back is the first 200 words or so of each item – not the whole content. Those 200 words or so are processed by the enterprise search application to produce the cleverly arranged and prioritised results list from all the various sources. So if you want your organisation's contribution to come high on the list, get the first 200 words right.

Much more difficult is the question of cross-references. Google, for example, is widely believed to use the quantity, and quality, of hypertext links as part of the results list weightings. The theory also suggests that links to your material are more important than links from your material. So, if some highly reputable organisation links to your paper on uncertainty in contracts that will help it get to the top of the list.

Fixing titles

Good editors of websites, internal knowledge resources, and online sites, carefully craft accurate, full titles for each discrete chunk of content. They do not accept, uncritically, the titles given by the authors, whether they are too short ("lease") or off the point. They craft them for one purpose – getting found on websites – and they pay very serious attention to them, since the title is the most important thing. Two tips:

- Boring is good: a title up to fifteen words long, getting in as many of the vital words as you can, is what you should create. "The legality of regulatory practice on bankers' bonuses" is a good title. So is "Shell and core office leases, VAT and corporation tax traps, drafting notes". Both titles tell your readers (and the search engine) what the pieces are about. They describe the text. What more do your potential readers/users want?
- Clever titles don't work. "Gun-smoke and dogs' tails: a counter-blast to Heisenberg's Principle" is a really silly way to title your article on some subject such as uncertainty in contracts. It's fashionable to use cute titles – so let everyone else do it. They get lost; you are found. (This author has been guilty of some wretchedly cute titles in his time. Alas, I have had to grow up).

Fixing the first 200 words

Remember – this top part of the document is critical real estate; it is the equivalent of the shop window for a shop. There are two elements here, and both are vital. They are:

- Checking for, and removing, initial document clutter;
- Writing the very best abstract you possibly can; with a third caveat, which is counter-intuitive, but usually true:
- Don't waste time on keywords fields (use "keywords" differently).

Initial document clutter is insidious, but you need to get to grips with it. If you do not know XML/HTML, get someone from IT to help you. Open the vital articles – the ones you have had loaded to the website in the hope that they will be found by new clients or users – in raw ASCII format (eg use "Notepad"). After all, that's what the search engines do. Now look at the first few lines. You may be dismayed to find that most of what you see, in each of your articles, is nothing to do with title and abstract, but a whole lot of rubbish, perhaps even about the organisation that runs your website. There can also be other material there which may be needed, but should not be in the first few lines.

Somehow, you have to fix this and you may make yourself unpopular with the web site people in the process. Make sure the title and first paragraph of text (see below on this) are right up there in the earliest part of the document. Get the entries like: <!– This material is managed by Henrico's Super Website Management and Patented Content Development System: visit us at www.henricosisbest.com –> moved lower down or removed.

Do not take "no" for an answer, and do not believe people who say "that is OK, the search engine takes no notice of that". If they were right, why would anyone bother to put the initial document clutter there in the first place? Anyway, nobody knows the arcane search engine rules at any one time, and you know they cannot know. Get any internal document IDs and reference numbers moved lower down, which usually means the bottom, instead of the top. Your top lines are critical and you donot want a lot of meaningless numbers taking up space.

The easy bit for LIM readers is fixing the content of the abstract. Actually, it is the first paragraph: whether you call it an abstract, or summary, or it is just the opening paragraph, is not that important. Opinion varies on whether tagging the first paragraph "summary" helps

https://doi.org/10.1017/S1472669610000071 Published online by Cambridge University Press

or hinders. Use all your skills and do not leave it to the authors. Your classification techniques and skills will help you enormously here. Think how you would classify the material. Look up the synonyms (you do not know how your potential users will search, so help them out!) Now work all those vital words into the first paragraph.

You may get the best results if you adopt a catch-phrase style – two hundred years of case reporting cannot be wrong. It means that more of your words count. But do not overdo it. Search engines often include a test for connected, grammatical text and ignore, or downgrade, material outside the title which does not pass the test.

It is very doubtful if including a "keywords" tag in this document real estate will help. Because keywords fields (tags) were widely and heavily abused, the rumour is that Google et al now routinely ignore such tags. Stick to that first paragraph of actual text.

Exploiting hypertext links

This is where you are least likely to be successful, so this is a short section. If you have a contact in some hugely prestigious organisation, and can get lots of hypertext links to your website inserted in their material, well, worth a try. You could set up an interlocking circle of back-scratching cross-referees. But on the whole, the material you want your clients to find will not be readily referenced by others and will have limited references going outwards.

Nonetheless, there is no harm in adding hypertext links provided they work, and you make them to well-respected organisations. Links to free government websites will probably get you more Brownie points in the search engine algorithms than links to paid-for sites. But this is a grey area; that is why it is the last priority.

Keep the articles small

Split up long articles into separate sections. That way, instead of one title and one 200-word shop-window, you get three, or four. Use "Part 1" etc to show your users that there is more of the same, and provide the hypertext links between them. A suggested rule-of-thumb could be: Work on an outside limit of three thousand words, and normally go for smaller chunks.

Automatic classification

The purposes

Some of the reasons to use auto-classification are important, but not germane to legal information management. For example, the ability of auto-classification to identify document types (invoices, letters, contracts, parts lists, meeting agenda, etc) is clearly very useful in large corporate business processes, but this is rarely a big problem for legal organisations. Again, the use of auto-classification techniques to file large numbers of documents in a DMS, under various virtual folders, is invaluable if you are installing, changing or merging a DMS. But that is not usually the province of *LIM* readers. There are other uses as well. But here are three that are important to legal information.

First, classification by hand ought to be accurate, but in practice it can be inconsistent, and it is invariably expensive and slow. In any case, now that so many organisations use enterprise search in one form or another, material is being brought in from outside in far greater quantities than the internal content. Obviously the external material can't be classified (a tiny amount of it is classified by its creators, but usually using taxonomies rather different from those used for the internal material). Auto-classification can materially assist with all these points – if it works well.

Secondly, for most "search" purposes, the desired pattern is well established:

- The user types a simple term or phrase, in a "Google-style" box;
- The results come back organised and counted by topic, concept, etc.

The second part of this pattern is where auto-classification comes in. Since all the material being searched cannot possibly be classified by hand, then logically this second activity can only be accomplished by computers. Thus a much greater range of material can be analysed far more quickly than is possible by manual techniques.

Thirdly, large discovery projects present huge challenges to those charged with managing them. Auto-classification clearly has a major role in e-discovery applications. It is a specialised subject and beyond the scope of this article. If you want help on this, and do not know where to go, I can point you at good people who do this for a living. That is all I shall say specifically about e-discovery here, but it is important to get to grips with it, and what I say about reference sets below can be an absolutely vital tool for you.

Does automatic classification work?

If it is straight out of the box, then not very well. It takes some configuration work, by people who know what they are doing. Then, results that are 70 – 90% as good as the best manual classification work can be achieved – but applied over an enormously expanded range of materials. In many cases, results at this level require the use of "training sets" - ie document collections which have already been classified, which the software uses to learn how to auto-classify material.

At the end of this article I give two references to validation studies, whose text is available free online. My assumption is that this will be a starting point for you – after all, you are the people who are good at research.

https://doi.org/10.1017/S1472669610000071 Published online by Cambridge University Press

But be a bit wary, because most validation studies suffer from two constraints:

- They are produced by vendors, rather than independent researchers (most auto-classification development is done for commercial exploitation);
- They use training sets which are much larger in size than those available to most legal organisations (where, say 2,000 well-classified documents are available, rather than 10,000 or 30,000).

The problem

The theory is great. It goes like this. For the last 20 years, an army of earnest legal information professionals (that is you, by the way) have busily classified, according to the most stringent indexing disciplines, and using thesauri and taxonomies constructed according to the strictest principles, many thousands of documents. So now, those are the training set; the amazingly clever software just reads those documents and notes the classifications applied; then off it goes to look at unclassified documents, compares them to the classified ones and adds in the appropriate classifications automatically.

That hard work, and the clever software, should now have come of age. What was a joke in 1989, and the province of geeks in 1999, is perfectly able to be done today. Everyone is a gainer, nobody is a loser.

But there is a catch. For a number of reasons, many of which we thrashed out in a memorable pair of sessions at the BIALL Conference in Harrogate in 2005, taxonomies and classification are not exactly favourite recipients of budgetary cash in legal organisations. We need not recapitulate the whole story, but to summarise:

- Over the last decade, lots of people became amateur taxonomists, and created over-large, poorly constructed classification schemes unsuited to the small legal know-how collections to which they were to be applied. A depressing amount of that work, and the classification work on documents based on it, was therefore wasted.
- The vast majority of lawyers will not use an overtly taxonomy-based search (though some will use a taxonomy-based drill-down browse facility more readily).

Implementing automatic classification

Some organisations did not fall into this trap. They have created well-classified document collections, which can be used as reference collections. The advice of this article to these organisations is to make use of that work and leverage it. Talk to the auto-classification people – usually, that means the search integrators/enterprise search providers. There is a big range of price, and a corresponding range of performance; but you can get a huge advantage in accuracy, across a range of material you could never possibly classify by hand, in this way. You have an important head start over your competitors, because you have created the reference set which matches all sorts of content to the concepts and terms used in your organisation.

To those who have not done a classification project, or whose classification projects have yielded at best equivocal, at worst useless results, this article advocates adopting SEO techniques. Here are the reasons:

- Internal search integration ("enterprise search") applications use the top part of the documents, as discussed in detail above;
- It is smart to adopt the techniques which will work whether the material you are working on is destined for the web, internal use or client use – and the distinctions are becoming increasingly blurred;
- You can quietly adopt simple, accurate classification terms, in the wording of the vital first paragraph, without all the emotion and committees associated with poorly executed taxonomy projects – this was discussed above, under how to make the most of the first 200 words;
- Remember to include synonyms, because this will help the software, and the users, alike. The old mantras of go and stop terms get quite blurred nowadays; don't resist this, embrace the change;
- Auto-classification software can be told to weight the first paragraph for concepts (it almost certainly will anyway).

Conclusion

You can tweak internal engines – a weighting here, a rule there – but you can not tweak Google, so you have to learn about SEO if you are to operate usefully to your organisations in the web world. Auto-classification is not perfect, but it is rapidly becoming an essential productivity tool which cannot be ignored. It has specific uses in legal organisations and you should be leading the charge. You have two really helpful trends on your side:

- Preparing a reference set for auto-classification can either use existing classified documents (typically know-how material) if you have them, or SEO techniques; either way, you face forwards, not backwards;
- Simplification is the watchword: whatever you do with manual classification must be to simplify, not=complicate, the terms and your work, and to bring it back where it belongs – in information services.

https://doi.org/10.1017/S1472669610000071 Published online by Cambridge University Press

References

Al-Safadi, Lilac A. E., 2009: Auto Classification for Search Intelligence. *World Academy of Science, Engineering and Technology* 49, 849–852 http://www.waset.org/journals/waset/v49/v49-150.pdf

Eysenbach, G. and Ch. Kohler, 2003: What is the prevalence of health-related searches on the World Wide Web? Qualitative and quantitative analysis of search engine queries on the Internet. *AMIA Annual Symposium Proceedings* 2003: 225–229. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1480194/

Biography

Derek Sturdy is a knowledge management consultant at Tikit Ltd. Not everyone at Tikit would agree with everything in this article, which represents his views, not the company's. He also runs the Galilee Project, a not-for-profit research operation collating and verifying past climatic information for use in validating climate models. derek.sturdy@tikit.com.

Legal Information Management, 10 (2010), pp. 28–33

doi:10.1017/S1472669610000277

Indexing of Free, Web-based Electronic Resources

Abstract: The internet provides access to a huge amount of information, and most people experience problems with information overload rather than scarcity. Glenda Browne explains how indexing provides a way of increasing retrieval of relevant information from the content available. Manual, book-style indexes can be created for websites and individual web documents such as online books. Keyword metadata is a crucial behind the scenes aid to improved search engine functioning, and categorisation, social bookmarking and automated indexing also play a part.
Keywords: indexing; internet; metadata; classification; automatic indexing

Introduction

Despite the success of search engines, human-created indexing remains important to enhance access to the most relevant information for searchers' needs. This article covers a range of access methods, focussing mainly on the areas in which human input is most significant, including website indexes (back-of-book style A–Z indexes to websites and web documents), metadata creation, social bookmarking, and classification, followed by a brief discussion of automated indexing. Searches for 'website indexing' also retrieve information about webwide search engines, but they are not discussed here.

1. *A to Z indexes*

Website indexing became important in the 1990s, as indexers, librarians and web managers experimented with different approaches for making the information they were providing on the internet more accessible. The tools for creating A–Z indexes have changed over time, from simple HTML coding to HTML Indexer and other