

Is searching the best way to retrieve legal documents?

by Trygve Harvold

Abstract

Legal texts have a rich structure and a large number of links which can be utilized in retrieving documents. This paper is based on a numerical study of the link structure in approximately 200.000 documents in the Lovdata database. The hypertext structure is analyzed and it is suggested that it should be possible to navigate the database on the basis of indexes and links. Analysis of the use of Lovdata also indicates that utilizing chronological and alphabetical indexes and the hyperstructure of links might in many cases be a more efficient and user-friendly way of finding documents than the traditional search.

1. Retrieving legal documents

The Lovdata Foundation has been running a legal information system since 1983. Today Lovdata has thousands of users including all the larger law firms and government bodies in Norway. A major challenge to Lovdata in all these years has been to provide a user friendly interface.

Searching has been the traditional way of retrieving documents. This is true for legal information systems, as well as for Internet as a whole. However, it is not a big secret that searching presents serious problems for many users.

There are two general types of searches:

1. The document you want to find is known to you and you know how to find it, that is you know a specific character string (or word) which occurs in the document. Retrieving a court case on the basis of its identification number is an example of this.
2. The documents you want to find are unknown to you or you do not know any specific character strings (or words) in the documents. This is normally the case when you want to find documents which concerns one or more subjects in which you are interested.

For type 1 situations search engines work just fine - there are really no problems whatsoever. For type 2 situations search engines have serious problems, especially if the document collections get large. The main, and obvious, reason for this is that the same subject can be described in various ways and with various words depending on the author and the type of document. If you don't already know the documents you are trying to find, locating them by way of a search can therefore be very difficult.

In the case of the Internet this turned out to be such a big problem that Google became a huge success by putting much more emphasis on the ranking of documents than on searching in the traditional sense.

Ranking of documents in databases involves sorting the documents according to certain criteria. Thus ranking in a sense is very similar to searching. What Google realized was the following: Ranking the retrieved documents according to their similarity with the query failed, due to the same reason that the search itself normally failed, namely because there was not enough information in the query to go on in the first place. Most users make queries of just one or a few words, and this does not provide enough information for the system.

Google therefore does not rank documents according to their similarity with the query, but according to their significance in the database. More specifically the so-called PageRank ranks a document according to the number of other documents which links to it ¹⁾. A Google search then essentially consists of two steps: first the words in the query are expanded automatically by other similar words and then the retrieved documents are sorted by use of the PageRank in such a way that all the “junk” documents, which presumably nobody links to, are shifted to the bottom of the list.

Google has become a huge success precisely because you get to the “serious” documents first. However, Google does not necessarily give you the documents you really need or want.

2. An alternative to searching

In legal information systems there are in fact alternatives to searching²⁾. Documents such as statutes, regulations and court cases have a rich and homogeneous structure which makes it possible to establish chronological, alphabetical and systematic indexes. In addition this approach utilizes the hyperstructure of the documents which has been established by expert knowledge

A section in a statute will often be the natural starting point for a legal search. Most professional users are able to locate a statute on the basis of a chronological, alphabetical or systematic index. From the relevant section, they can then use the hyperlink structure to access the other documents in the system.

To see better why this is possible, we shall look closer at the structure of links in legal source documents.

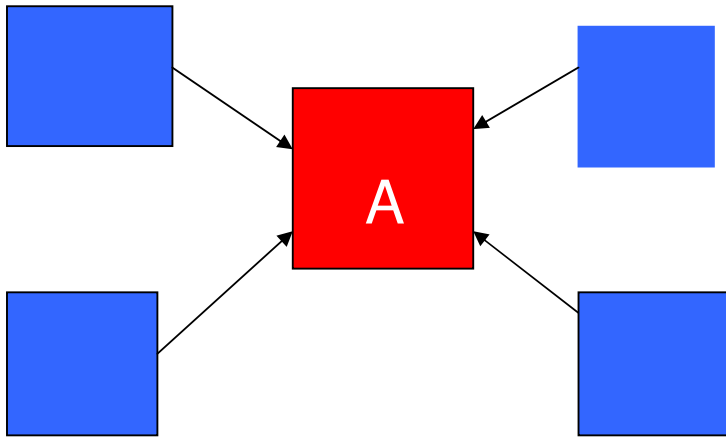
3. The hyper structure

In this paper we shall refer to the following three types of links:

- Link: natural link in document A to document B
- Backlink: constructed link in document B to document A which has a natural link to B
- Fatlink: link to one or more documents, ie for example all the backlinks in a document.

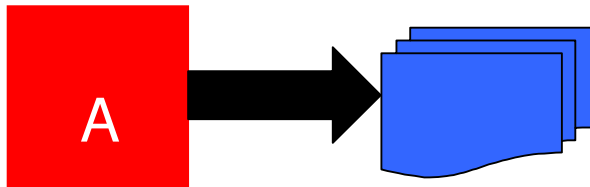
Links and backlinks are thus mirror versions of each other, and a fatlink, as used in this paper, is the collection of all backlinks in a document. A fatlink is represented as a button in the Lovdata system.

Links to document A

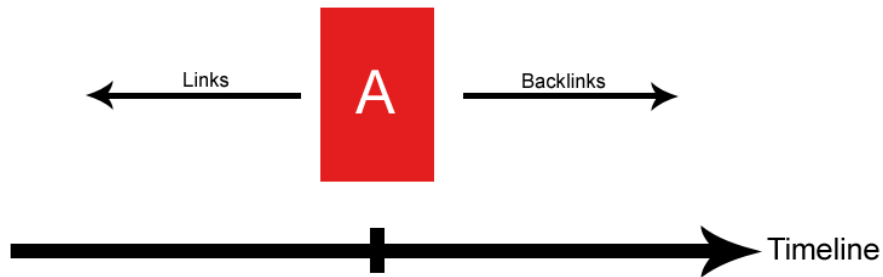


Every link to document A results in a corresponding backlink from document A.

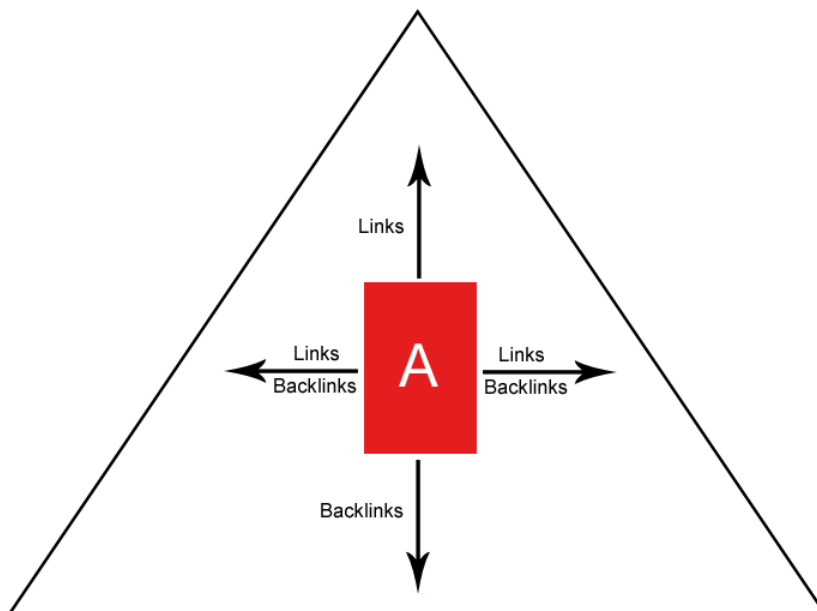
Fatlink (button) from document A back to the documents which link to A.



We note that links which are part of the original text of document A always go to older documents than A. Likewise backlinks from document A always go to newer documents than A.



If we consider document A as part of the legal hierarchy of documents, we note the following. The links from A are likely to be to documents higher up or at the same level in the hierarchy. The reason for this is that legal documents normally cite documents with higher or equal authority. For the same reason the backlinks will most likely point to documents of lower or at the same level in the hierarchy.



This figure show that it is at least theoretical possible to reach any document in a database by way of links or backlinks. Whether this is a practical possibility depends on the frequency, distribution and density of especially the lawlinks in the documents. We would like to have a large number of lawlinks, evenly distributed among the documents and a high density of lawlinks for each document type.

4. The frequency, distribution and density of links in the Lovdata system

The below numbers are based on a numerical study of the links and backlinks in most documents in the Lovdata system.³⁾

4.1. Frequency of links

Legal sources make up a hierarchy according to their importance in legal decision making. Statutes and supreme court cases are at the top of the hierarchy, and legal literature and first instance court cases somewhere near the bottom.

It is natural to assume that the number of backlinks in a document says something about the importance of the document in this hierarchy. This is so because the number of backlinks in document A is identical to the number of links to document A. In table 1 the document types are thus ranked according to the number of backlinks for each document type.

Table 1.

Document type	No of backlinks
Statutes	1.118.803
Supreme court cases	194.934
Preparatory works	120.810
National regulations	100.839
Other decisions and opinions ^{a)}	79.504
Appellate court cases	31.822
EU-directives/regulations	28.539
Local regulations	17.958
European Court of Human Rights ^{b)}	7.161
Published legal papers ^{c)}	4.880
EU-cases	4.602
First instance court cases	2.310

- a) Especially the insurance boards have a lot of citations to their own cases, e.g AKN (The Board for Reduced Compensation) (15787) and FSN (The Insurance Agreements Board) (48747)
- b) Citations to ECHR first started to appear in the early nineties
- c) This number is too low because not all links are identified

We see that this ranking probably corresponds fairly well to most lawyers' "feeling" of what the hierarchy should look like. The reason the decisions by the ECHR are ranked fairly low is that "human rights" in practice only applies to a few areas of law, which necessarily limits the number of backlinks. The same applies to EU-law.

Table 2 ranks the different document types according to their likelihood of having a fatlink. The table reflects the degree to which the different document types function as a legal source by being cited by other documents. The table mirrors table 1 to a certain extent, but is more detailed with respect to document types.

Again we see statutes and supreme court cases on top.

Table 2.

Document type	No of doc	No of documents with no fatlinks	Fatlink ratio
Statutes	777	7	0,99
Supreme court - civil cases	15395	2299	0,85
Supreme court - penal cases	26543	7846	0,70
Supreme court - appeal cases	17959	6435	0,64
Administrative opinions	17901	8465	0,53
National regulations	3888	1874	0,52
Preparatory works	10594	6443	0,39
Appellate court - civil cases	32783	20379	0,38
Appellate court - penal cases	14710	9089	0,38
Published legal papers	3349	2121	0,37
Parliamentary resolutions	225	154	0,32
First instance court - civil cases	8265	6017	0,27
First instance court - penal cases	2542	2043	0,20
Instructions	206	170	0,17
Local regulations	5009	4160	0,17
Regulatory delegations	1075	950	0,12
Special courts ⁴⁾	20584	18029	0,12
Circulars	1064	948	0,11
Administrative decisions	19617	18602	0,05

a) Fatlink ratio: $\frac{\text{number of documents with fatlinks}}{\text{total number of documents}}$

We note that administrative opinions have a fatlink ratio of 0.53, while administrative decisions have a fatlink ratio of only 0.05. This means that opinions tend to cite each other to a much larger extent than decisions. I have not analyzed the specific reasons behind this.

4.2. Distribution of links

We would expect a high and fairly even distribution of lawlinks across document types

The reason for this is that all other legal sources derive their “authority” from statutes, and should therefore, according to legal custom, cite the relevant statutes. Of course not all legal issues are covered by statutes, and in documents dealing with these issues, we would expect a lower distribution of lawlinks.

Table 3.

Document type	Distribution ratio of lawlinks
Appellate court - penal cases	1,00
First instance court - penal cases	1,00
Local regulations	1,00
Parliamentary resolutions	1,00
National regulations	0,99
Statutes	0,99
Supreme court - penal cases	0,99
Appellate court civil cases	0,98
Regulatory delegations	0,98
Supreme court - appeal cases	0,98
Administrative decisions	0,89
First instance court – civil cases	0,89
Instructions	0,88
Supreme court - civil cases	0,85
Published legal papers	0,82
Special courts ⁴⁾	0,81
Administrative opinions	0,76
Circulars	0,53
Preparatory works	0,50

a) Distribution ratio of lawlinks = $\frac{\text{number of documents with lawlinks}}{\text{total number of documents}}$

We note that the ratio is very high for most document types.

The ratio should be 1.00 for all databases with penal decisions, since these decisions have to have a legal basis in statutes. However, some of the penal cases by the Supreme Court are very old and lack a formal citation to the relevant statute.

We note that *Preparatory works* have a very low distribution ratio. This database includes documents which are not specifically law-related, for example some of the *Official Norwegian Reports* (NOUer) are not law-related. This also is the case for some parliamentary proposals and so on.

4.3. Density of lawlinks

With respect to court cases, administrative decisions and opinions, we would expect a higher density of lawlinks at the top of the legal hierarchy. The reason for this is that towards the bottom of the hierarchy, the legal issues are less likely to be law-related, and other legal sources are more likely to be taken into account.

Table 4.

Document type	Density ratio of lawlinks
Supreme court - penal cases	0,85
Appellate court - penal cases	0,82
Supreme court - appeal cases	0,81
First instance court - penal cases	0,80
Administrative decisions	0,76
Special courts ⁴⁾	0,76
Appellate court – civil cases	0,74
First instance court - civil cases	0,70
Supreme court - civil cases	0,64
Published legal papers	0,42
Administrative opinions	0,34

a) Density of lawlinks = $\frac{\text{number of lawlinks}}{\text{total number of links}}$

The density ratio illustrates the relative importance of statutes in the different document types. Penal cases mostly cite statutes (other citations are presumably to other supreme court cases and to preparatory works). The documents at the bottom of the list (administrative opinions, legal literature, civil cases and so on) also cite at lot of other types of legal sources – the reason of course being that they deal with issues which are less dependent on statutes for their resolution.

5. How Lovdata works

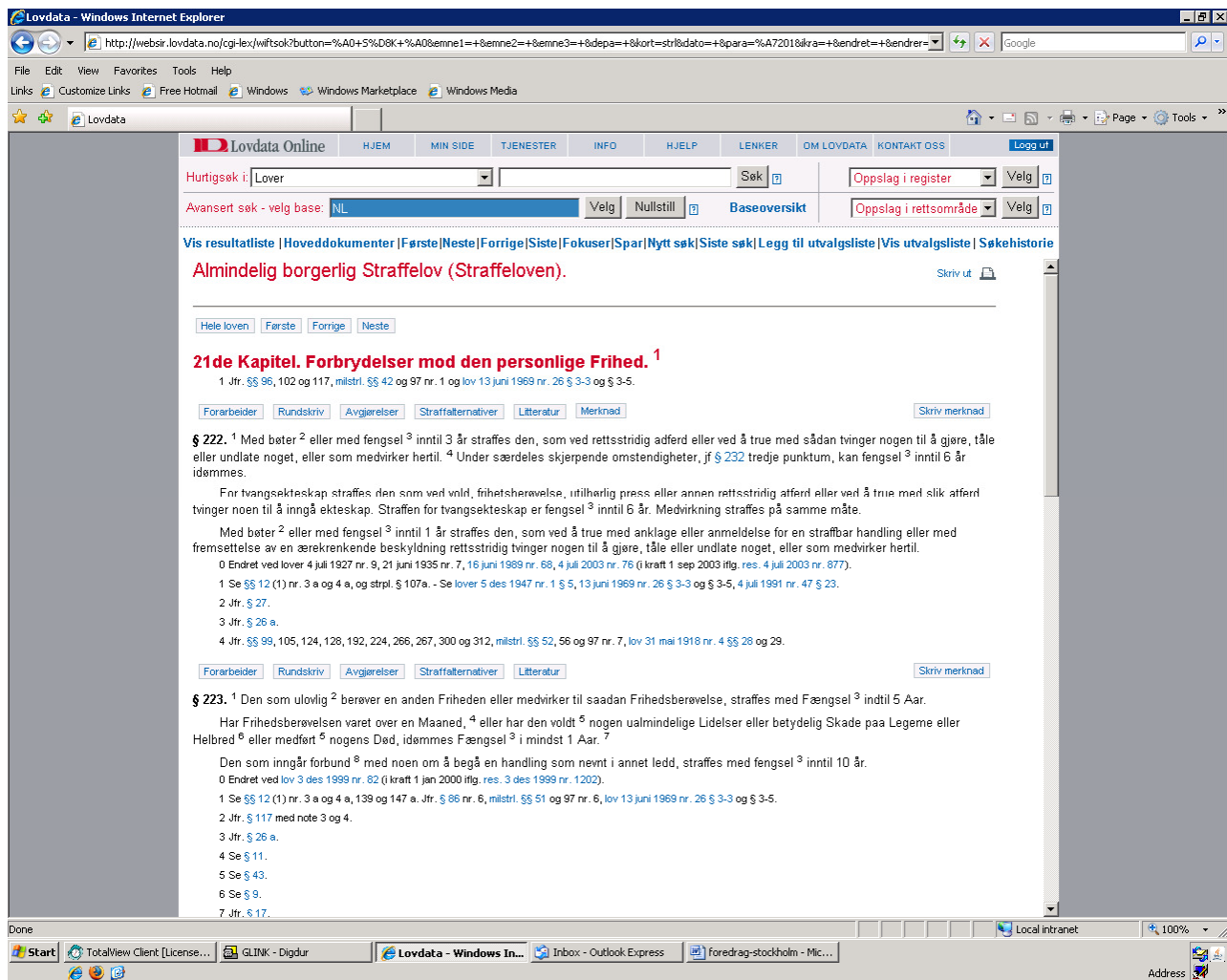
Users of Lovdata have four different ways of locating any given document:

- Quick search
- Advanced search
- Use of chronological, alphabetical or systematic indexes ⁵⁾
- Use of an index made up of legal subjects.

Both the *Quick search* and *Advanced search* are powerful search functions which can transform and expand the query based on both the specified field and the specified words.

Once you have reached a document, you can navigate further by way of the buttons.

Buttons appear normally at the top of a document. Buttons are fatlinks which make it possible to link to all other documents which link to the document. In the case of statutes and regulations there may also be buttons for each section, see below.



There are nine different types of buttons. The button types associated with the different document types are as follows:

Statutes	Regulations	Cases / literature	Preparatory works
Historical versions	Historical versions	Circulars	Document history
Amendments	Amendments	Cases	
Regulations	Preparatory works	Literature	
Preparatory works	Circulars	Commentaries	
Circulars	Cases		
Cases	Literature		
Literature	Commentaries		
Commentaries			

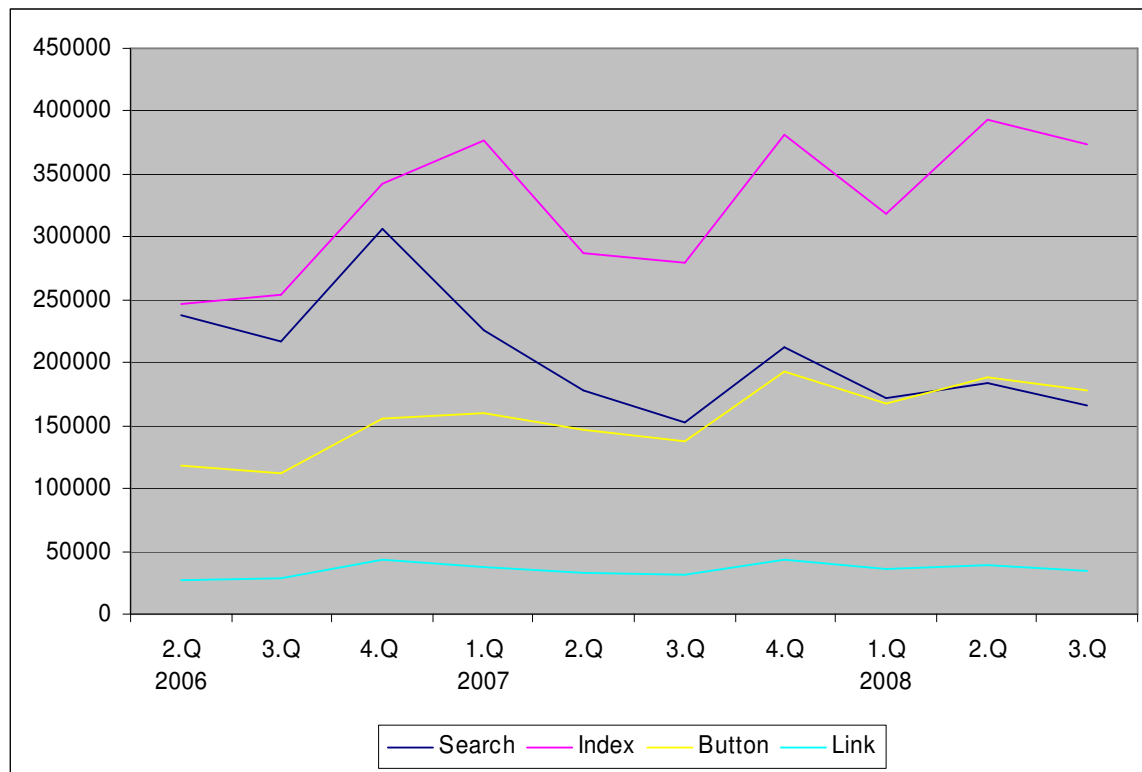
The content behind the button **Commentaries** are written by the users themselves. There are three types of commentaries:

- Personal commentaries
- Commentaries for the firm
- Commentaries for all

Any user can write a commentary and decide who will be able to read it by selecting one of the above three categories. The *Commentaries* button will only materialize for a given document if there exists a commentary which the user can read.

6. Statistics on the use of Lovdata

Indexes, buttons and links have been part of the Lovdata system for a long time. However, at the start of 2007 the system was given a new design which attempted to exploit these functions in a more systematic way. The graph below illustrates that in the last couple of years there has been a shift away from searching to the use of indexes and buttons. The use of normal links have remained fairly stable.



Even if the graph illustrates a trend towards a more frequent use of indexes, and vice versa for searching, it does not necessarily support the assumption that indexes are used more often than searching in today's system. The reason we cannot draw such a conclusion is that it normally takes three or four clicks to locate a document by way of the indexes.

7. Conclusion

While searching is a necessary and powerful tool, it may not always be the most user-friendly way of locating documents in a legal information system. In this paper we have shown how the rich structure and numerous links of legal documents allow for the construction of indexes, buttons and links which makes it possible for users to navigate the system without searching. User statistics from Lovdata show that users often prefer this alternative way of navigation in situations where it is possible and practical.

8. Notes

1. Actually the PageRank (named after Lawrence Page) is a little more complicated since the links themselves are weighted according to the importance of the document in which the link is situated. See Sergey Brin and Lawrence Page, *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, Computer Networks and ISDN Systems Volume 30 1998.
2. See for example Kirill Miazine: *Rettskilder og hypertextstruktur: om alternative grensesnitt til rettslige informasjonssystemer*, Complex 6 2006, Oslo
3. The numerical study is based on a total number of 220.787 documents in the Lovdata database (pr. 08.08.2008)
4. ARD = Labour Court, LODA = cases from Lov&Data , JSO, JSR = Land Consolidation Courts, NAD = labor cases), NDS = Nordic maritime cases, TRR = Social Security Court.
5. In fact there are also other indexes as for example indexes for ministries, municipalities, protected geographical areas, and so on.

