Theme:

Title:

# Identification and Classification of A/E/C Web Sites and Pages

Author(s): Ye Chen and Robert Amor

Institution(s): University of Auckland, Auckland, New Zealand

E-mail(s): trebor@cs.auckland.ac.nz

Abstract: *Current search engines are not well suited to serving the needs of A/E/C professionals. The general ones do not know about the vocabulary of the domain (e.g., so 'window' is a meaningless word) or rely on human classification (which severely limits the percentage of sites which are indexed). Domain specific databases and hot lists tend to be the only other option. While these have very good information they reflect a very small proportion of what is on the web.*

*This paper looks at a system for automated classification of web sites and pages in the A/E/C domain. In particular, we concentrate on web sites and pages in New Zealand, and use the common classification system for the New Zealand construction industry (CBI). For this particular problem it is clear that no single approach to classifying web information gives a perfect answer. We therefore combine several approaches for automated classification, including:*

- *Identifying web sites that are already classified by other Internet portals and mapping these classifications to the CBI classification system.*
- *Extracting keywords from web pages and sites and then finding the relationships between the extracted keywords and topics in the CBI classification system.*
- *Using link analysis to find related web pages on a certain topic in the CBI classification system.*

*When an A/E/C professional searches with our system we determine metrics for each approach above, and find the best combination of approaches to determine a classification and hence the resultant web sites and pages.*

*This paper describes the components of the search engine which has been created and provides an analysis of the classification approaches.*

Keywords: *Search engine; Classification; Internet*

## Introduction

While general search engines (e.g., AltaVista, Google, AskJeeves, etc) provide an extremely useful service for general queries they lack the structure and domain specific knowledge to make them a practical tool for work within a discipline. It is also clear that their concentration on covering the whole Internet is unfeasible and their coverage of particular topics and countries can be patchy. Previous research (Lawrence and Giles 1999) determined that the top search engine only covered 16% of the Internet's publicly indexable web pages, and the top 11 search engines together managed approximately 42%. A previous investigation of search engines and web site hotlists in the construction domain (Amor et al 2000) highlighted the fact that approximately 75% of the content of any one site was unique to that site. This means that the view of the information in construction which a user gets from such a service is very dependent upon which service they choose to use.

In the European Community funded CONNET project (Turk and Amor 2000) an initial attempt was made to develop a search engine for the construction industry that was more comprehensive than existing services. This Signposts system (2000) provided an AltaVista-like service to a large number of mainly English language construction sites. However, analysis of this system made it clear that it still covered only a very small percentage of web sites relevant to construction and the pure word search capability (i.e., searching for the word(s) entered by the user) was not always identifying the most appropriate web sites.

In this project we have taken a more analytical look at how web sites and web page content can be identified and used to build a search engine which better identifies relevant information for its users in the construction domain.

### *The premise for a search engine providing highly relevant information*

The premise of this work is that by using a single common classification system as the basis for identifying all web site and web page content it is possible to deliver highly targeted information to the users of the search engine. To test this premise it is necessary to be able to accurately classify a web site and all of its pages to a specific classification system. It is also necessary to be able to map a user's query to codes within the classification system, or to get them to choose classification codes to form their query.

In the remainder of this paper we will assume that a user either chooses a classification code for their query, or that the problem of mapping a user query to a classification code is a subset of the problem of classifying a web page. Hence, we will concentrate on approaches to classifying web site and web page content.

### *Methods of classifying content*

Current approaches to automatically extracting information from web content does not get us to a classification code. These methods include:

- Word search: searching for words only allows a match when there is a direct correlation between the terms used by a user and words which exist in a particular page or site. This can be made more powerful through the use of synonyms and a thesaurus (for example WordNET 2002).
- Keyword extraction: this technique extracts words from a document which encapsulate the key information in that document (Kea 2002). However, the keywords that are extracted can only come from the document content (i.e., it is not a controlled vocabulary).
- Link analysis: this technique establishes measures of similarity of concept by comparing links to and from a web page to those of a known concept. This approach provides a very good match between concepts, but it requires the web pages to have some history. For example, new and pertinent information on a topic will not feature highly until links are made to it by other sites.

Current approaches to hand classifying web site information do provide a high quality classification for the content that has been seen. However, the classifications which are currently applied to web sites are ones which have been developed by search engine companies and not the ones in use by a particular industry. It is also clear that current human efforts can not keep up with the amount of material which is being created on the web. The ODP initiative (ODP 2002) has classified over three million web sites through the effort of 47,000 volunteer editors. However, when each of the big search engines index over two billion web pages it is clear that there is a huge gap between what humans can cope with and what the search engines need to work with. We also note that the human classifications are just for a web site as a whole and are not applied to individual pages within each site.

### *Approaches to automated classification*

In this project we explore the accuracy of indexing web pages through a range of techniques. In order to measure the effectiveness of each technique we have created a set of hand classified web sites and pages against which the automatic approaches can be compared. The techniques that we have examined include:

- Term comparison: In this approach we compare words from the codes of a classification system with words in a web site or a web page. Where there is a close match between the words from a classification code and the content of a web site or web page then that classification code is applied to the web site or web page. In this approach we also examine the use of synonyms and thesaurus expansion in the matching.
- Keyword comparison: In this approach the keywords for a web site or web page are extracted and compared to the words from the codes of a classification system. Where there is a close match between the words from a classification code and the keyword content of a web site or

web page then that classification code is applied to the web site or web page. In this approach we also examine the use of synonyms and thesaurus expansion in the matching.

- External classification mapping: In this approach we utilise the classification that has been assigned to a web site or web page by another search engine or catalogue (e.g., Yellow Pages 2002). We examine automated mapping of codes from one classification system to the base system used for the search engine, as well as utilising an expert's specification of the mapping between two classification systems.
- Link analysis: While link analysis can not classify a site directly we can use its ability to identify similarities between web pages to specify a classification. To do this we start with a set of expert classified web pages and use link analysis to match with other web pages inside the search engine. Where the measure of similarity is high, the same classification can be applied to the web page in the search engine.

## Preliminary evaluation

In the following section we look at the preliminary results from our analysis of the range of techniques to classify information content.

### *Method of evaluation*

We used the Co-ordinated Building Information classification (CBI 1998), the common classification system of the New Zealand building and construction industry, as the core classification system of the search engine. CBI is a four-level hierarchical classification system. Each code in CBI has a descriptive tag and a long description.

The evaluation to another classification system was done by examining the construction part of the New Zealand Yellow Pages directory (Yellow Pages 2002), which indexes the greatest number of construction web sites in New Zealand. The Yellow Pages classification system is also hierarchical and has at most three levels with a total of 229 classification codes. Each code has a short descriptive tag but no long description.

To accomplish full-text searching we used the full-text search system in Microsoft's SQL Server. This full-text search can look for words or phrases and their inflectional forms in a character text, and returns a ranking score in terms of relevance.

To provide a base mapping for comparison we asked experts in construction to manually map Yellow Pages codes to the CBI classification system. This resulted in 101 human-specified mappings. It is important to note that a large amount of the classified content in Yellow Pages is not able to be directly mapped to the CBI classification system. We also chose 50 web sites at random from the database of known New Zealand web sites (totalling just over 2000) and had these manually classified with CBI codes. The classifier provided keywords for these 50 random web sites.

### *Effectiveness of keywords extraction and synonyms*

To test the effectiveness of keyword extraction we put the descriptive tag and long description of each CBI code together as searchable text, and full-text searched each Yellow Pages descriptive tag against this text. In order to measure whether keyword extraction and synonyms are useful, we used Kea to extract keywords from the searchable text and used WordNet to return synonyms for each keyword. We also used CBI's descriptive tags as searchable text and expanded the CBI descriptive tag with synonyms after removing stop words (e.g., a, an, the, etc.).

Table 1 provides an example of the expansions which are obtained by this process for a sample of CBI codes.

| | |
|---|---|
| *CBI code* | 38-31 |
| *CBI name* | Timber floors |

| | |
|---|---|
| *CBI description* | Timber floors, constructed of solid or manufactured timber flooring, fixed to framing. Acoustic sheeting, fixed to framing to form substrate to timber flooring. Timber strip flooring, with edges tongued and grooved or butt jointed, fixed to floor framing, either face nailed or secret nailed. Flooring of manufactured timber board, fixed to framing. Prefabricated timber flooring/decking, fixed to framing. Sports floors of timber strip flooring, tongued and grooved, end matched and secret nailed to battens, with or without a plywood or acoustic underlay, set on resilient pads and fixed to the concrete substrate over a vapour barrier. Sports floors of parquet or thin timber strips, stuck to a plywood or acoustic subfloor which is set on resilient pads and fixed to the concrete substrate over a vapour barrier. Flooring of timber forming stair treads, walkways and catwalks, fixed to bearers. Flooring of manufactured timber boards faced with thin strips, thin parquet or plastic laminate, fixed to framing. Sound insulating packers and supports. Thermal insulation and/or vapour barriers fixed at the same time as the flooring. Proprietary or non-proprietary supporting clips, fixed to timber or concrete subfloor. Sanding, sealing and otherwise preparing surfaces and applying protective coatings and decorative finishes off site. Preparing surfaces to receive protective coatings and decorative finishes on site. Forming openings and cutting holes as required to suit sub-contractor and General Contractor responsibility. Non-fixed work. Related work in other work sections: On site finishing (66). Fully supported timber flooring (62-31) |
| *CBI name and synonyms* | Timber, Lumber, forest, woodland, timbre, quality floors, floor, flooring, level, storey, story |
| *Keywords from CBI description* | timber, flooring, timer flooring, fixed to framing, timber strips, substrate, manufactured timber, sports floors, floors of timber, tongued and grooved |
| *Keywords and synonyms from CBI description* | timber, lumber, forest, woodland, timbre, quality<br>flooring, floor, shock, stun, ball over, blow out of the water<br>timber flooring, lumber, timber, forest, woodland, timberland<br>fixed to framing, repair, mend, fix, bushel, doctor<br>timber strips, lumber, timber, forest, woodland, timberland<br>substrate, substratum<br>manufactured timber, manufacture, fabricate, construct, cook up, make up<br>sports floors, sport, athletics, sportsman, sportswoman, mutant<br>floors of timber, floor, flooring, level, storey, story<br>tongued and grooved, tongue, tongued |
| *Classifications and scores from a full-text search of Yellow Pages tag 'Timber flooring' against keywords and synonyms from CBI description* | **CBI code** **Score**<br>38-31 268<br>38-3 255<br>75-45 190<br>38-33 141<br>38 115<br>… |

***Table 1**. Example of CBI code expansions and search results*

Figure 1 shows the number of classifications returned from full-text searching of all the options described above, compared with the hand mapping. This search is applied across all classification codes in the Yellow Pages system. We let Kea return at most 10 keywords and WordNet return at most 5 synonyms to expand the keywords. The choice of these parameters are discussed later in this paper. The full-text search returns a list of CBI codes ordered by ranking scores for each Yellow Pages code submitted. We examine the top 10 through to the top 50 of these CBI codes and see whether they contain a CBI code exactly the same as the one in the hand mappings. If there is, we treat it as a correct classification.

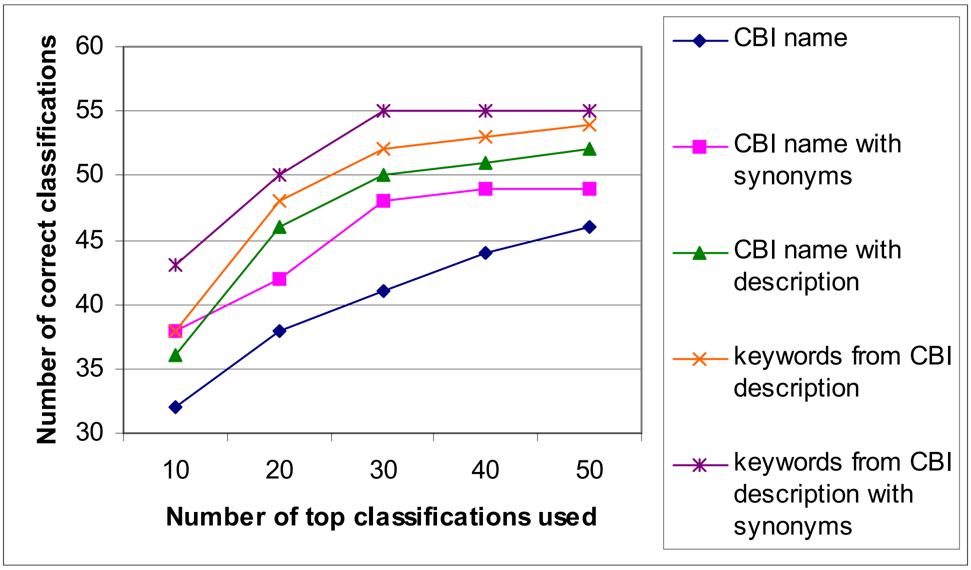


***Figure 1**. Effectiveness of keywords and synonyms*

Figure 1 shows that using keywords from the CBI description performs better than using the whole description. The CBI tag terms with synonyms perform better than the CBI tag terms alone, which demonstrates that synonyms help the system find words or phrases with a similar meaning to Yellow Pages codes. The best results are produced by keywords from the description extended with synonyms.

*Effect of keywords*

We investigated how the number of keywords extracted from a description affected the performance of classification. We let Kea return different numbers of keywords from CBI descriptions, and full-text searched Yellow Pages tags against different numbers of these keywords expanded with synonyms.

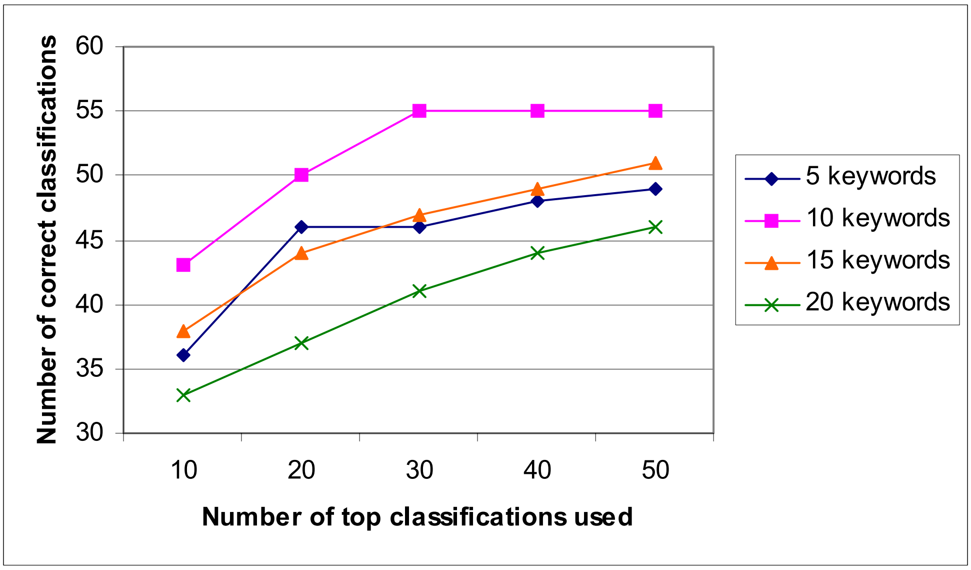


***Figure 2**. Effect of number of keywords extracted*

Figure 2 shows that when 10 keywords are extracted from the CBI description, the system finds the most number of correct rankings comparing with the hand mapping. When 5 keywords are extracted, a smaller number of correct rankings were found, which shows that 5 keywords does not adequately represent the content of the searchable text. When the number of keywords extracted increases above 10 the performance drops. We can conclude that as more keywords were extracted more irrelevant words were introduced.

*Effect of synonyms*

In order to measure how the number of synonyms extracted affects the system's performance, we expanded keywords from the CBI description with different numbers of synonyms returned by WordNet. We used 5 and 10 synonyms and compared this with the performance of keywords from the description without synonyms.

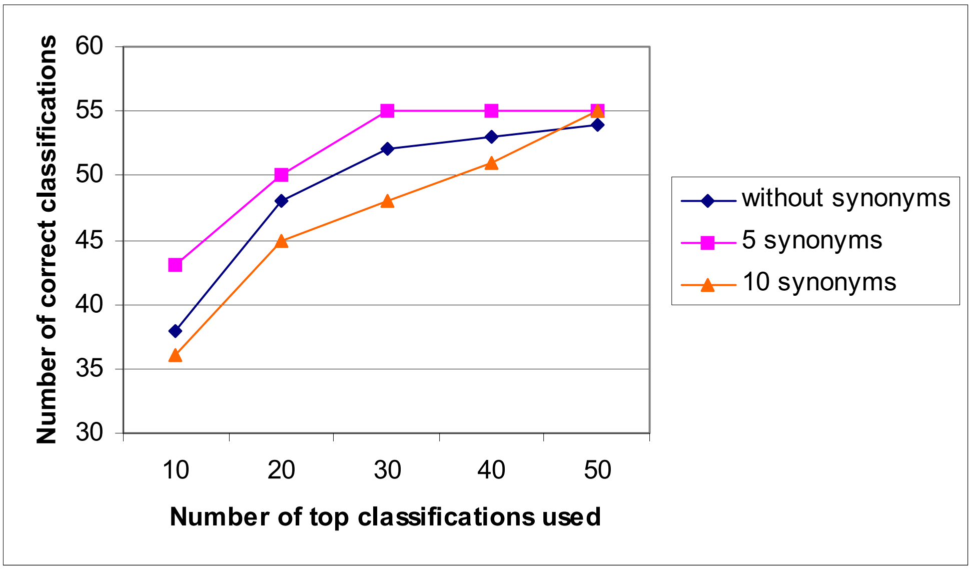


***Figure 3****. Effect of number of synonyms used*

As seen in Figure 3, when 5 synonyms are used to expand the keywords, the system returns the highest number of correct rankings. It shows that keywords expanded with a small number of synonyms work better than without synonyms, while a large number of synonyms introduces irrelevant terms and causes a drop in the performance of system.

*Effect of ranking number and ranking scores*

Using the list of ordered rankings returned by SQL Server's full-text search for each Yellow Pages code, we examined the effect this has on finding the best CBI code. We investigated how many rankings should be used and how the ranking scores are useful. We assumed that the top rankings with higher ranking scores are more important than rankings with lower scores, so we removed rankings with scores less than a certain percentage of the top ranking score.

This analysis shows that using the top 10 rankings finds the best CBI code and returns the greatest number of correct mappings to the classification system. We also found that removing returned rankings with scores less than 55% of the top ranking score helps improve the performance of the system.

We were also interested in the ratio between correct and incorrect codes which are returned by this method. Figure 4 illustrates the ratio of correct mappings compared with the total number of mappings returned. In this case the top 15 rankings return the largest ratio of correct mappings, while the performance of the top 10 rankings is lower, which indicates that the top 10 rankings return more incorrect mappings.

Considering the number and ratio of correct mappings, we believe that using the top 15 rankings to get the CBI code is a better choice, when rankings with scores less than 55% of the top ranking score are removed. At this point 83 Yellow Pages codes are automatically mapped to CBI codes and 29 of these are correct mappings, which accounts for 35% of the mappings.

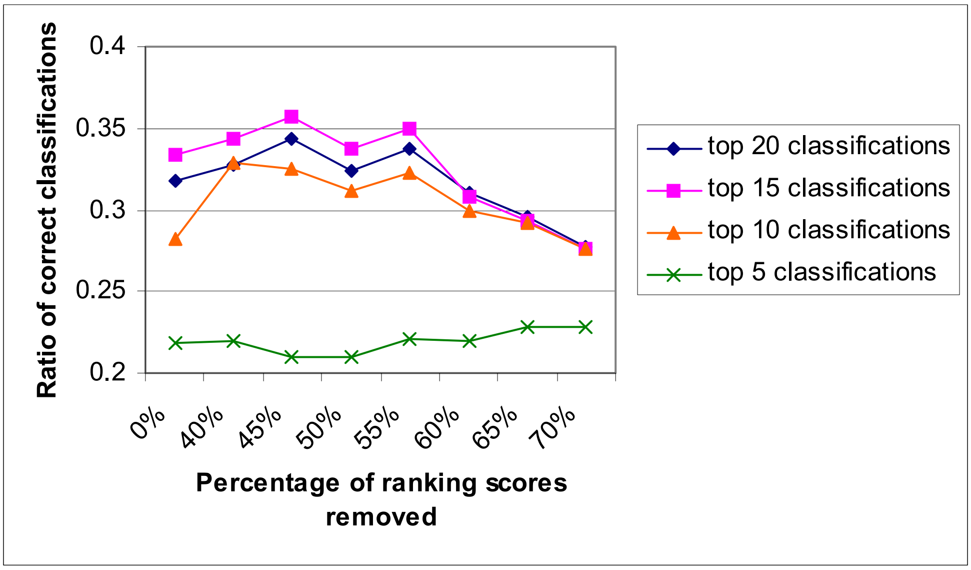


***Figure 4****. Ratio of correct mappings with effect of ranking number of ranking scores*

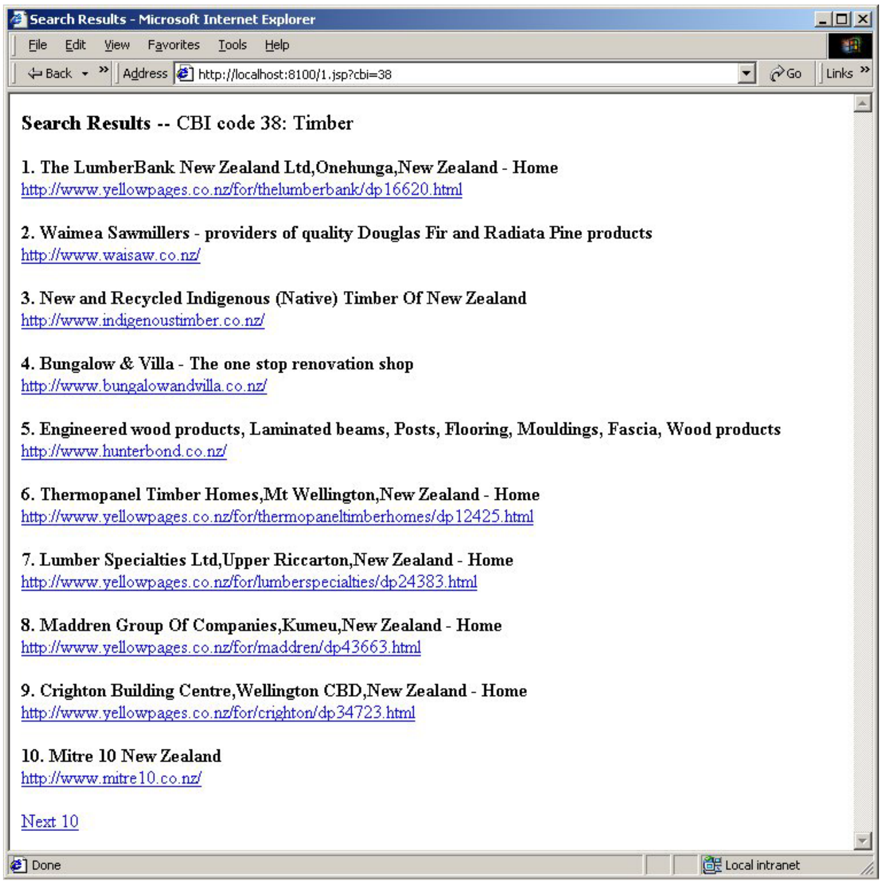
### Preliminary search engine

Based upon the results reported above we have created an initial search engine for New Zealand's construction related web sites. Our search engine queries web sites or web pages by a given CBI code or CBI name. Figure 5 shows the result of searching for CBI code 38 (Timber).

### Conclusions and future work

This paper presents initial work on developing a search engine based upon a construction industry classification system. As part of the development we have analysed the impact of different approaches to the automated mapping of information to classification codes. We have shown that the choice of expansion method and controls on the amount of information returned by each method has a considerable impact on the final accuracy of classification code selection by the system. It is clear that the use of keywords and synonyms provides a more accurate mapping than using the original text. It is also clear that using a restricted set of keywords and synonyms reduces the amount of superfluous information and increases the accuracy of selection of classification codes. By restricting the use of ranked information returned by the full-text search system utilised in this project we can also increase the accuracy of the selection of the classification code. However, we do acknowledge that the automated classification is not highly accurate, with only around a third of classifications being correct in relation to those identified by a human.

The next stage in this work ties the appropriately configured automated classification of information into the user interface for the search engine. This will be the real test of the efficacy of the approach in comparison to existing search engines. The premise is that even though the accuracy of classifications being applied is not high, we should still get better search results than when using a search engine with no knowledge of the construction domain. To test this premise we will test our search engine's results against those of Google and AltaVista when constrained to a similar set of web pages (i.e., New Zealand construction pages).


Search Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Address
Go
Links
Done
Local intranet

Address http://localhost:8100/1.jsp?cbi=38

**Search Results** -- CBI code 38: Timber

1. The LumberBank New Zealand Ltd,Onehunga,New Zealand - Home
http://www.yellowpages.co.nz/for/thelumberbank/dp16620.html

2. Waimea Sawmillers - providers of quality Douglas Fir and Radiata Pine products
http://www.waisaw.co.nz/

3. New and Recycled Indigenous (Native) Timber Of New Zealand
http://www.indigenoustimber.co.nz/

4. Bungalow & Villa - The one stop renovation shop
http://www.bungalowandvilla.co.nz/

5. Engineered wood products, Laminated beams, Posts, Flooring, Mouldings, Fascia, Wood products
http://www.hunterbond.co.nz/

6. Thermopanel Timber Homes,Mt Wellington,New Zealand - Home
http://www.yellowpages.co.nz/for/thermopaneltimberhomes/dp12425.html

7. Lumber Specialties Ltd,Upper Riccarton,New Zealand - Home
http://www.yellowpages.co.nz/for/lumberspecialties/dp24383.html

8. Maddren Group Of Companies,Kumeu,New Zealand - Home
http://www.yellowpages.co.nz/for/maddren/dp43663.html

9. Crighton Building Centre,Wellington CBD,New Zealand - Home
http://www.yellowpages.co.nz/for/crighton/dp34723.html

10. Mitre 10 New Zealand
http://www.mitre10.co.nz/

Next 10

***Figure 5****. Screen snapshot of the search engine results*

## References

Amor, R., Marsh, R. and Hutchison, A. (2000) Electronic News Service for the European Construction Industry, Proceedings of Construction Information Technology 2000, Reykjavik, Iceland, 28-30 June, pp. 52-62.

CBI (1998) CBI: Co-ordinated Building Information, http://www.masterspec.co.nz/CBI-1.htm, last accessed 1/4/2002.

Google (2002) Google search engine, http://www.google.com/, last accessed 1/4/2002.

Kea (2002) Kea automatic keyphrase extraction, http://www.nzdl.org/kea/, last accessed 1/4/2002.

Lawrence, S. and Giles, C.L. (1999). Accessibility of Information on the Web, Nature, No. 6740, July 8, pp.107-109.

ODP (2002) ODP: Open Directory Project, http://www.dmoz.org/, last accessed 1/4/2002.

Signposts (2000) Signposts: the construction industry search engine, http://online.bre.co.uk/signposts/, last accessed 1/4/2002.

Turk, Z. and Amor, R. (2000). Architectural foundations of a construction information network, International Journal of Construction Information Technology, 7(2), pp. 85-97.

WordNET (2002). WordNET: a lexical database for the English language, http://www.cogsci.princeton.edu/~wn/, last accessed 1/4/2002.

Yellow Pages (2002) New Zealand Yellow Pages Business Phone Directory, http://www.yellowpages.co.nz/, last accessed 1/4/2002.