Automated classification of A/E/C web content

R. Amor & K. Xu
Department of Computer Science, University of Auckland, Auckland, New Zealand

ABSTRACT: The amount of useful information available on the web for A/E/C professionals increases inexorably. Numerous search engines allow users to identify potentially useful information in this vast resource, though the majority of these systems work purely on the search terms entered by the user. This means that the web pages which are found are often not as relevant to the user's needs as would be expected. What is returned is certainly far from the promise of the semantic web where the properties of the content can be readily ascertained. To help address this issue the authors adapt the Latent Semantic Indexing algorithm to enable web pages and sites to be automatically matched to codes in a classification system. This paper discusses the issues involved in developing such a system for A/E/C as well as measuring the results in comparison to the general search engines currently available to professionals.

1 INTRODUCTION

In this paper the adaptation of a standard information retrieval technique, namely latent semantic indexing, is examined for a domain specific search engine. The premise behind this approach is that it is possible to accurately identify classification codes related to the content of the web page or web site. If content can be accurately classified then a user searching for content in a particular area (e.g. by specifying a classification code) will be presented only with highly relevant web information.

The reason that we attempt to classify to a standard classification code is that these are used and understood by the vast majority of professionals within the A/E/C industries. Because a classification code has a well described scope it is likely to be understood similarly by professionals from many disciplines. Therefore, a system that can accurately retrieve information associated with a classification code is one which can be tied to many processes within the A/E/C profession where information is associated with these codes.

This paper describes the ongoing development of the LSI-based search engine. It concentrates particularly on the testing of the resultant search engine in terms of the precision of the classification of construction industry web pages to a construction industry classification system (by comparison with an expert's determination of correct classifications). It also provides an analysis of the developed search engine's search result accuracy by comparison with the results returned by other major search engines in current use (e.g. Google and Yahoo) on the same query formulation.

1.1 Previous work

The field of information retrieval has grown rapidly since the 1940's (Chu 2003), though the majority of the useful approaches to retrieval were developed in the early years of the field. With the birth of the Internet information retrieval has gained further prominence and the major search engines (e.g. Google and Yahoo) are used by a large percentage of people in the western world. Though there are large search engines which are used frequently by those in business they are usually based on fairly simple retrieval algorithms (to achieve their speed requirements) and by serving all domains for all people will often throw up irrelevant results to any specific query.

Domain specific search engines are being investigated and in the A/E/C industry there have been previous approaches to establishing such search systems. In the EU-funded CONNET project the Signposts system was developed (Signposts 2000, Turk & Amor 2000) based on Boolean search of keywords, though utilizing web page structure to help rank results. Further developments of this system utilized the HITS algorithm and mappings from pre-classified web page repositories (Chen & Amor 2002) to help improve the precision of returned web pages. However, none of these systems were signifi-

cantly better than the commercial search engines of that time.

Research has also been undertaken to classify research papers within the A/E/C field through automated means (Turk & Cerovšek 2003), though again the authors report discrepancies in the results returned by the automated process.

1.2 Latent Semantic Indexing

To allow web content to be classified more accurately the LSI (Latent Semantic Indexing) algorithm (Deerwester et al. 1990) has been examined. This algorithm looks at all the words in a document which are relevant to the domain, and uses counts of these relevant words to help determine similarity to another item (which may be a search term, or another document).

The basic idea behind the algorithm is that a set of keywords (and their synonyms) is selected to cover all the required concepts in a domain. The words and phrases in each web page, or web site, are then indexed against every one of the previously selected keywords. If a matrix column, representing a web page or site, is then compared against a similar matrix which has been computed for all of the terms in a classification system then it is possible to identify classification codes which are most closely related to a particular web page or site (by calculating the cosine of the angle between the two vectors that comprise these matrix columns).

Therefore, as web pages are gathered from the Internet they are compared to the matrix representing a particular classification system and for each classification code a similarity measure is assigned. When a user of the search engine requests information, by specifying a classification code, the system can then return a ranked list of web pages and sites which are determined to be the most close to that code. If a user searches by specifying keywords and phrases the LSI algorithm can be applied to their search term to identify the most relevant classification codes, and then identify the most relevant web pages or web sites.

2 SEARCH ENGINE APPROACH

To test the applicability of LSI for classification of web content for A/E/C it was necessary to establish domain specific terms and vocabularies to drive the LSI algorithm. This section describes where this information was derived from and how it was applied for our search engine.

2.1 A/E/C terms

A success factor for the LSI algorithm is the choice of suitable domain specific terms which cover all concepts in the area in which information is to be retrieved. For the A/E/C domain there are very few digital resources which profess to cover the vocabulary of the professions. However, the LexiCon (Woestenenk 2002) has been developed over many years to try and capture terms used in construction, and the relationships between them, based on the wide variety of construction specific classification systems in use across the world.

The LexiCon database was used as the seed for the set of terms required for the LSI algorithm. Mining the LexiCon database provided over 15,000 terms covering the whole A/E/C domain. However, not all terms were unique, as some represented synonyms of other terms in the system. To enable similar concepts to be captured as a single term in the LSI algorithm each of the LexiCon terms was expanded with known synonyms and each of these sets of terms were treated as a single concept. As part of this thesaurus expansion the terms were stemmed to reduce the susceptibility of the system to problems of singular, plural, and other forms that terms commonly take in natural language text.

2.2 A/E/C classification system

This search engine has been developed to allow the use of the developers preferred classification system. However, as each web page is classified against all terms in the classification system there is quite a cost to introducing a new classification system into the search engine. As this search engine is developed to serve an audience of New Zealand's A/E/C professionals the standard classification system for the country was utilized.

The CBI (1999, Coordinated Building Information) has over 1000 classification codes in a four level decimal classification structure. The CBI classification system provides a textual description for each of the codes and it is this piece of text, along with the title of each code, which is used to populate the term matrix derived from LexiCon.

While the descriptive text associated with each code is not extensive (usually about a paragraph) it does provide the system with a much larger set of words than a plain code title with which to populate the term matrix as required in LSI.

2.3 Populating the page repository

To provide a source set of web pages to examine the system's performance, just over 14,000 web pages were gathered drawn from an existing database of web sites know to be relevant to A/E/C in New Zealand (Chen & Amor 2002). In this search engine a conscious decision was made to distinguish between web sites and web pages. A web site comprises all web pages found below a particular URL. For example where http://www.sopers.co.nz/ is identified

as the entry point for a web site, all web pages identified by links from this page, but still part of the Sopers domain, are tagged as part of this web site.

With this distinction in place a user can retrieve web information either at the specific web page level for detailed and closely matching information related to their search, or at the web site level, which is closer to a view of the company associated with information on the particular search area.

2.4 A user interface

A user of the search engine is provided with two methods of searching the repository, either through the classification system, or by keyword search.

2.4.1 *Classification-based search*

A user can search by providing a known CBI code (e.g. 38-3 for Timber floors, stairs and covers) or by navigating the tree representing the CBI classification system to identify a code they wish to utilize.

Once a code is specified the web page and web site indexes are utilized to identify pages and sites most closely associated with that code. Results of the search are presented to the user in a ranked order as shown in Figure 1.

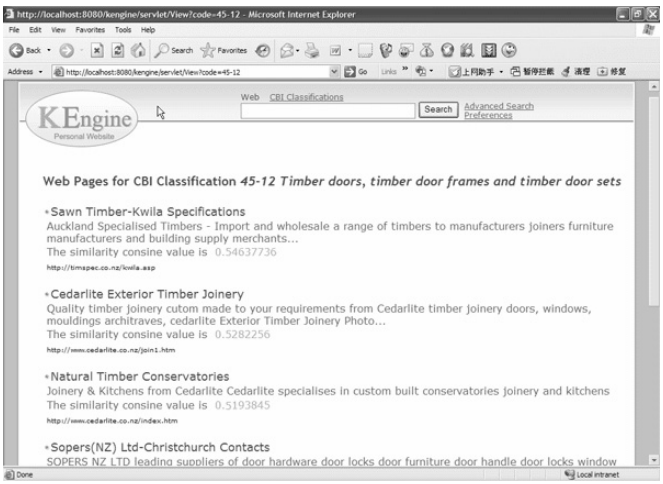

Figure 1. Result screen from classification code search.

2.4.2 *Keyword-based search*

A user can also search by providing a set of keywords. In this mode the search process has two steps. In the first step the user's keywords are matched to the CBI classification system. This is achieved by using the LSI algorithm to determine from the set of CBI codes which is closest to the entered keywords. The resultant list of ranked CBI codes is proffered to the user for them to select their desired code, or to navigate the CBI classification hierarchy from one of those codes to identify their preferred code. Once the code is identified the search proceeds as for the classification-based search.

3 RELEVANCE OF RESULTS

Determining the accuracy of this new search engine proves difficult. In the information retrieval domain there are measures to be used for this task (Chu 2003), however, many of these measures (e.g. recall) require knowledge of the set of relevant information items which were not retrieved. As can be imagined, with any collection of useful size, this figure is almost impossible to ascertain (it would require every web page to be hand classified).

3.1 A measure of precision

For the analysis of this system we have chosen to use precision as out main measure of accuracy. Precision is a measure of the ratio of relevant documents retrieved versus the total number of retrieved documents. For our testing we have further restricted this by analyzing the relevance of the first twenty documents which are retrieved by search engines.

3.2 The testing framework

Our test suite comprises a set of twenty CBI classification codes. Ten of these codes represent more general topics in the CBI system (i.e. those at level two in the classification system) and the other ten represent more specific topics in the CBI system (i.e. those at level three in the classification system).

Testing against a search engine proceeded by examining the top twenty results returned by the search engine. Each of these results was hand classified as 'relevant' or 'not relevant' to the particular classification code. To perform our calculation of the precision we also recorded the total number of pages returned (as this can be less than twenty).

Searches were carried out against the new LSI-based search engine as well as through Google and Yahoo, both of which can be restricted to New Zealand sites.

3.3 Test results for the new search engine

For the new search engine this test was run both for web pages and web sites to determine what affect the larger set of information (i.e. web site) had on the accuracy of the returned results. Results from web sites and pages are shown in Figure 2.

Query topics	KEngine Web sites			KEngine Web pages		
	Related/20	Total	Relevance	Related/20	Total	Relevance
Demolition (21)	2	7	0.29	9	10	0.90
Masonry (33)	3	4	0.75	5	9	0.56
Timber (38)	15	58	0.75	19	1287	0.95
Doors, windows and roof lights (45)	14	71	0.70	17	401	0.85
Glazing (46)	15	32	0.75	16	75	0.80
Ceilings (53)	8	9	0.89	10	73	0.50
Carpeting (65)	9	21	0.45	11	208	0.55
Pools (84)	7	12	0.58	13	101	0.65
Concrete (31)	6	6	1.00	14	64	0.70
Ventilation and air-conditioning (76)	17	45	0.85	18	156	0.90
Natural energy (75-2)	6	7	0.86	11	201	0.55
Fencing and walling (83-3)	7	13	0.54	12	124	0.60
Space heating steam and hot water (75-4)	5	11	0.45	7	101	0.35
Aluminium doors, windows, and roof lights (45-2)	3	14	0.21	7	34	0.35
Thermal insulation (47-1)	2	3	0.67	18	23	0.90
Trowelledand sprayed coatings (61-1)	1	5	0.20	6	14	0.43
Timber floors, stairs and covers (38-3)	7	13	0.54	18	42	0.90
Exhaust systems (76-1)	1	3	0.33	4	10	0.40
Monitoring systems (78-5)	1	2	0.50	2	4	0.50
Pools, spas, saunas, and showers (84-1)	4	5	0.80	19	47	0.95
Average relevance			0.61			0.66
Average relevance (general topics)			0.70			0.74
Average relevance (specific topics)			0.51			0.59

Figure 2. Precision of results for web sites and pages

As will be noted from these results there is a significant level of variability in the precision of the searches. Of great interest is the fact that the precision is higher for general classification codes than it is for specific classification codes. This result was unexpected as the fact that a more specific vocabulary is used for specific classifications had suggested that this would lead to a better match than over more general terms. There is also a slight difference (0.61 versus 0.66) between the precision of retrieval of web sites versus web pages. This is less surprising as many web sites (i.e. companies) deal with many products and services and hence have a wider spread of terms to represent their complete site.

3.4 Test results for Google and Yahoo

To perform the same test in Google (2005) and Yahoo (2005) we accessed the country specific versions of these search engines (i.e. restricting the search to New Zealand web pages). Formulating equivalent queries for each classification code is of course not possible as the three search engines use very different search algorithms. The approach undertaken for this project was to use the classification code titles as the search term for both Google and Yahoo (e.g. Timber floors stairs and covers). In both cases there was no enforced grouping or sequencing of the words in the terms, though in some cases this would improve the results (e.g. 'air conditioning' as a sequence of words rather than 'air' and 'conditioning' as separate search terms). Results for Google and Yahoo are shown in Figure 3.

These results show the same trends as in the LSI-based search engine. Google appears to offer slightly higher precision than Yahoo, especially in the specific topics. In relation to the LSI-based search engine these results are remarkably similar, practically equivalent within the margins of error.

4 CONCLUSIONS AND FUTURE WORK

In this project a new search engine was developed specifically for the A/E/C profession within New Zealand. The aim of this system was to allow searches based on the standard classification system (CBI) utilized in New Zealand. To this extent the project is successful, allowing a user to search for web sites and web pages directly from a CBI classification code.

However, to be useful to the A/E/C profession in New Zealand the search engine should perform more accurately than the alternatives that people commonly use. The LSI-based search engine has been tested against two of the most popular general search engines (Google and Yahoo) and disappointingly it is found that the precision of the results returned by the new system is no different from that of other more generic search engines. Though these generic search engines do not allow searches purely on classification code (however the mapping to a search term is trivial).

Query topics	Google			Yahoo		
	Web pages			Web pages		
	Related/20	Total	Relevance	Related/20	Total	Relevance
Demolition (21)	15	4020	0.75	13	2820	0.65
Masonry (33)	18	2760	0.90	15	4950	0.75
Timber (38)	18	22100	0.90	17	16100	0.85
Doors, windows and roof lights (45)	11	108000	0.55	14	117000	0.70
Glazing (46)	14	2180	0.70	16	2840	0.80
Ceilings (53)	19	2640	0.95	15	2950	0.75
Carpeting (65)	10	423	0.50	11	654	0.55
Pools (84)	6	18200	0.30	8	23200	0.40
Concrete (31)	15	21000	0.75	14	21800	0.70
Ventilation and air-conditioning (76)	18	15100	0.90	18	13600	0.90
Natural energy (75-2)	12	169	0.60	10	115	0.50
Fencing and walling (83-3)	7	8250	0.35	8	7640	0.40
Space heating steam and hot water (75-4)	17	342	0.85	12	561	0.60
Aluminium doors, windows, and roof lights (45-2)	19	3010	0.95	18	2990	0.90
Thermal insulation (47-1)	19	643	0.95	16	678	0.80
Trowelledand sprayed coatings (61-1)	3	3	1.00	5	11	0.45
Timber floors, stairs and covers (38-3)	16	3050	0.80	11	654	0.55
Exhaust systems (76-1)	3	367	0.15	2	423	0.10
Monitoring systems (78-5)	6	1390	0.30	8	1250	0.40
Pools, spas, saunas, and showers (84-1)	7	25700	0.35	12	17865	0.60
Average relevance			0.68			0.62
Average relevance (general topics)			0.72			0.71
Average relevance (specific topics)			0.63			0.53

Figure 3. Precision measures for Google and Yahoo

All may not be lost however, the work presented to date is based on a very simple analysis of results and a more detailed analysis should be undertaken. Specifically a three or five point scale should be employed to mark the relevance of retrieved web pages, which is likely to paint a very different picture of the precision of each system.

It is also likely that amalgamating other retrieval techniques with LSI results (e.g. Boolean search terms based on a user query) will provide a more precise set of results for the user to work with.

REFERENCES

CBI 1999 CBI: Co-ordinated Building Information, http://www.masterspec.co.nz/CBI-1.htm, last accessed April 2005.

Chen, Y. & Amor, R. 2002. Identification and Classification of A/E/C Web Sites and Pages, *Proceedings of the CIB W78 Conference on Distributing Knowledge in Building, Aarhus, Denmark, 12-14 June*, 2, 37-44.

Chu, H. 2003. Information Representation and Retrieval in the Digital Age, *ASIST monograph series*.

Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W. & Harshman, R.A. 1990. Indexing by Latent Semantic Analysis, *Journal of the American Society of Information Science,* 41(6), 391-407.

Google 2005. http://www.google.co.nz/, last accessed April 2005.

Signposts 2000. Signposts: the construction industry search engine, http://www.connet.org/uk/signposts/, last accessed April 2005.

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

Turk, Ž. & Cerovšek, T. 2003. Mapping the W78 papers onto the construction informatics topic map, *Proceedings of the CIB W78's 20th International Conference on Construction IT, Construction IT Bridging the Distance, CIB Report 284, Waiheke Island, New Zealand, 23-25 April*, 423-432.

Woestenenk, K. 2002. The LexiCon: structuring semantics, *Proceedings of CIB W78 conference on Distributing Knowledge in Building, Aarhus, Denmark, 12-14 June*, 2, 241-247.

Yahoo 2005. http://au.yahoo.com/, last accessed April 2005.