Construction Management and Economics (1995) **13**, 411–416

KBS-CLASS: a neural network tool for automatic content recognition of building texts

JÖRGEN MODIN

KBS-Media Lab, Department of Structural Engineering, Lund University, Lund, Sweden

KBS-CLASS is a tool for automatic content recognition of building texts. It may be used for finding information in large or unstructured databases, for filtering news streams and as an aid for the classification of in-house-produced texts. There is likely to be a growing need for such tools as access to building knowledge in an electronic form becomes the key to an organization's success. The KBS-CLASS tool is based on neural network technology. It is trained with texts which have been indexed according to a classification system. The tool is then able to give a contents' description of texts it has not yet seen, with terms adopted from the classification system. The tool should be able to use any classification system, such as SfB, R-UDC, BSAB or the new ISO system under development. The current tool has been trained with texts from a building products database from AB Svensk Byggtjänst (the Swedish Building Centre) using the BSAB system. A case study is presented to show how the tool may be used in a future work situation. The tool's performance is discussed. Finally, future directions for further development of the tool and similar tools are suggested.

Keywords: Artificial neural network, information retrieval, building information.

Introduction

Communication of information in the construction industry and its storage and conversion will become more and more automated (Christiansson, 1993). This automation includes CAD interchange formats and Electronic Data Interchange (EDI) formats for trading (see, for example, Schlieper, 1992). However, this basic automation will also be accompanied by efforts to aid business at a more conceptual level. Powerful filtering mechanisms will be needed (Tetzeli, 1994). These filtering mechanisms must contain some kind of knowledge about what to retrieve. In response to these trends, KBS-CLASS has been developed as a tool for the automatic recognition of building texts. This paper describes the basis of the tool, in particular its construction which is based on the use of neural networks. A case study of KBS-CLASS is presented to indicate the future scope for its application and that of similar tools.

Several different approaches for computer-based text retrieval have been developed: those dealing with the way the computer performs the search and the way the user tells the computer what to look for. In terms of computer-derived searches, there are methods based on, for example, linguistics and rules (Cavazza and Zweigenbaum, 1992), statistics and knowledge bases (Croft, 1993; Jacobs, 1993) and even rhetorics (Sillince, 1992). Methods incorporating neural networks have also appeared (Lelu, 1991; Rose and Belew, 1991; Scholtes, 1993; Lamirel *et al.*, 1994). Neural networks have drawn much attention because of their ability to model complex relationships and form an important part of the tool to be described here, KBS-CLASS.

Internationally, efforts are being made to further the performance of information retrieval. Research institutes and companies have competed at the TREC (Text REtrieval Conference) with systems which automatically retrieve information from gigabytes of text. A wide variety of techniques were used and are described in the proceedings from the 1992 conference (Harman, 1993). In the US, a national institute for research in 'intelligent information retrieval' has been founded in Massachusetts (Croft, 1992).

User interface

Communication with the user is crucial in an information retrieval system. The most common user interface is one that handles Boolean queries. Boolean queries require the user to describe the required texts with words and Boolean operators (AND, OR, NOT, etc.). The user's words may differ from the vocabulary of the texts. The computer can help by searching for synonyms and

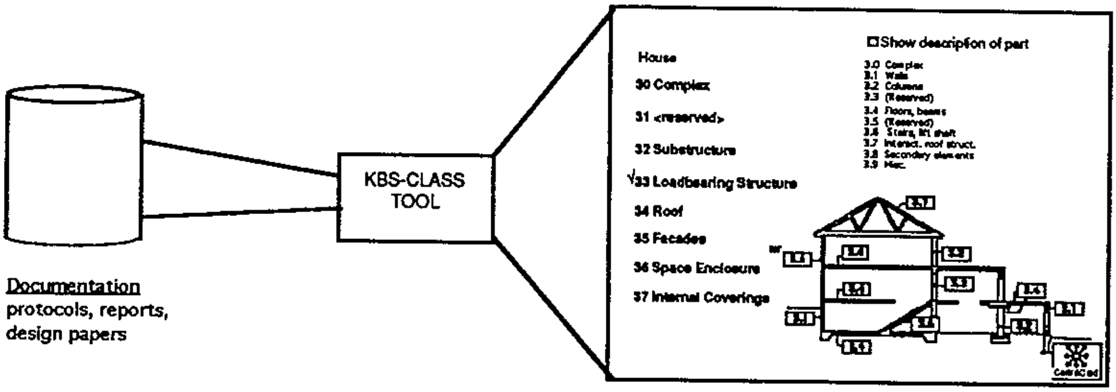


Figure 1 The KBS-CLASS tool lets the user retrieve unstructured information through the vocabulary of a classification system. The user interface to the right is from the Cube system (Christiansson and Modin, 1993)

related words. The user must know with some certainty what he or she is looking for, and it is difficult for people to use Boolean queries that contain several operators.

In natural language interfaces, the user states what he is looking for by using natural or near-natural language. The computer then analyses what the user has written and tries to find something that matches. The user might not know how to express exactly what he or she is looking for and even if he or she does it may be difficult for the computer to interpret the specification. To help the user specify what he is looking for, he could use ready-made examples of queries and ask for 'more of the same'. The computer returns new texts and the user could say 'more like that one' and the computer would use some kind of similarity measurement to deliver new texts. This method is called *relevance feedback* and is a way of relieving the user from forming their own natural language search queries from scratch. Relevance feedback is used, for example, in the WAIS system (Kahle, 1991).

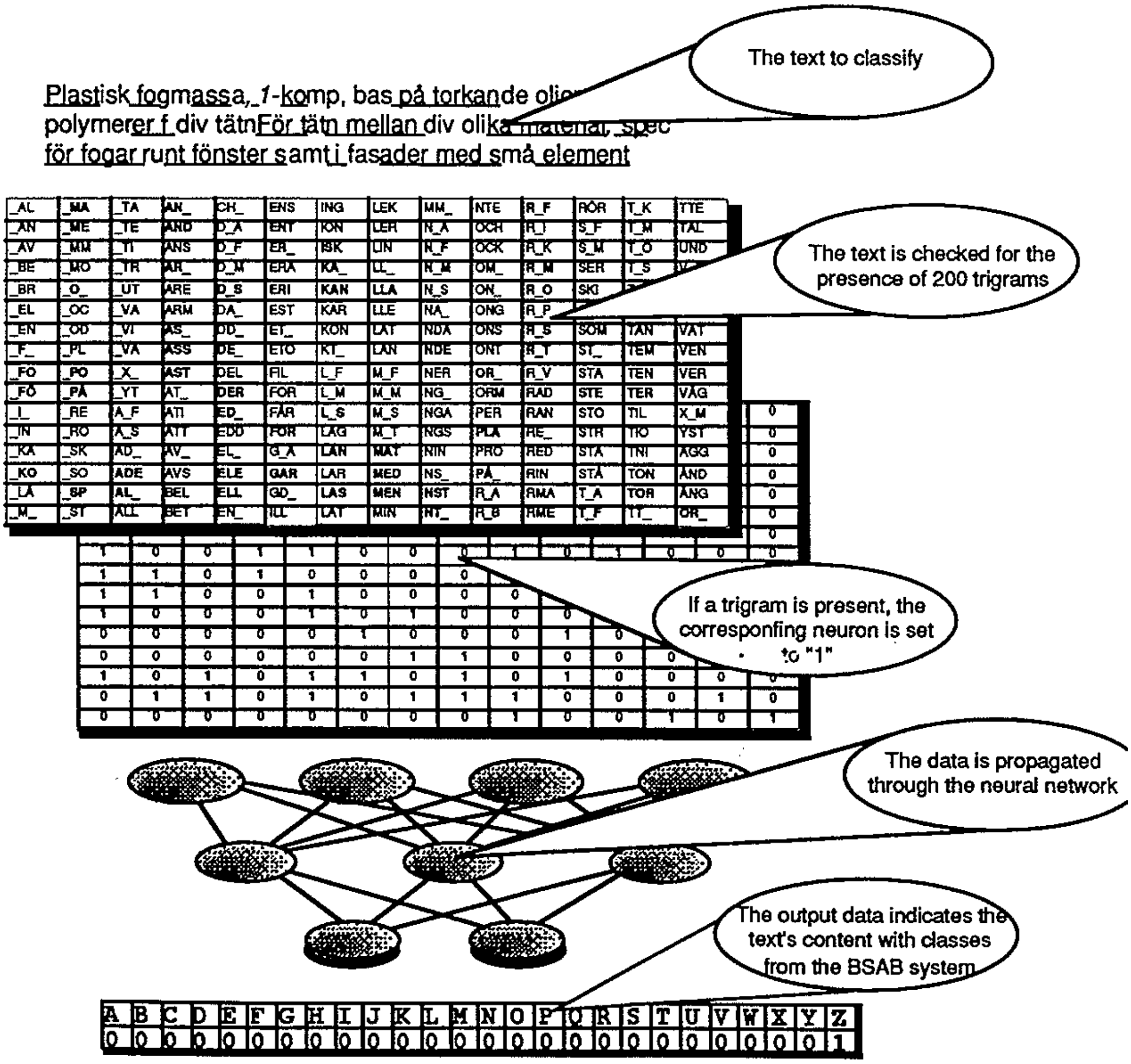
In KBS-CLASS, a classification system is used for the user interface. The classes can be seen as ready-made search queries that are easy for the user to understand. The user sees the classes on-screen and can point with the mouse to specify what he is looking for. He does not have to agree with the structure of the system, as long as he can use it fruitfully. This means that the classification system must have sufficient *semantic power* to provide effective queries and it must be a *good communicator* with little effort being required of the user to understand what a class means. In the Cube project (Christiansson and Modin, 1993), the BSAB system formed an important part of the retrieval interface. A test group of building staff found retrieval satisfactory (Berglund *et al.*, 1992). The communication of a classification system can be enhanced with illustrations as Fig. 1 shows.

Neural networks

The KBS-CLASS tool uses neural networks. In the context of computing, neural networks is a reference to a way of performing computations that are similar to certain aspects of how the human brain works. Specifically, one tries to compute in patterns rather than pre-defined symbols. A neural network on a computer consists of small simple processing units called neurons. These neurons are interconnected with numerous 'connections' (Lawrence, 1991) which are modified by the outer world. When the neural network receives a stimulus, it changes in response. The stimulus is a pattern of numerical data that is fed into a number of neurons. Example applications include diagnosing EKGs, forecasting sun-spots, forecasting financial market fluctuations or classifying texts. A neural network of the type presented in this article learns from examples presented to the network over and over again. Artificial neural networks is a dynamic field of research and there are few hard and fast rules about how they work or how they should be designed. For forecasting problems and for classification, a certain way of connecting the neurons and responding to stimulii has come to prevail: the *back-propagation network*. This network consists of neurons in layers, where each layer is massively connected to the next layer. This means that even a small network has a large number of connections. The KBS-CLASS tool has 34 300 connections between 385 neurons arranged in three layers. The back-propagation network is presented with input data to be classified and output data with the desired classification. The input data are propagated through the network in such a way that it is transformed into a classification. If the pattern, when it reaches the output, does not match the classification, the connections are modified slightly to nudge the network into performing the correct classification. After presenting many patterns – in the case of KBS-CLASS it is several thousands – the network learns to relate the input stimulus with the classification at the output. It is now ready to be presented with the input stimulus after which it can compute a sensible classification on its own.

KBS-CLASS

The input data to the KBS-CLASS neural network comes in the form of ones and zeros. A one indicates that a certain thing in the text is present and a zero indicates that it is absent. The things the KBS-CLASS tool looks for are *trigrams* (Scholtes, 1993). Trigrams are three-character chunks of text. The KBS-CLASS tool looks for 200 different trigrams (see Fig. 2). One could say that the presence of certain trigrams makes up a fingerprint or a signature for that particular text. The network is trained to relate this fingerprint to the classification of the text. Vast amounts of pre-classified texts are presented, and the network learns to generalize about the combinations of trigrams and the classifications. It learns that certain combinations of trigrams must be present for a certain classification and that the presence of other trigrams can invalidate that classification. Given enough input data and a soundly designed network, the tool starts to *see the big picture* and learns to generalize and to give reasonable classifications to combinations of trigrams it has not seen before.


The text to classify
Plastisk fogmassa, 1-komp, bas på torkande olje polymerer f div tätnFör tätn mellan div olika material, spec för fogar runt fönster samt i fasader med små element
The text is checked for the presence of 200 trigrams
If a trigram is present, the corresponfing neuron is set to "1"
The data is propagated through the neural network
The output data indicates the text's content with classes from the BSAB system
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1

_AL	_MA	_TA	AN_	CH_	ENS	ING	LEK	MM_	NTE	R_F	RÖR	T_K	TTE
_AN	_ME	_TE	AND	D_A	ENT	ION	LER	N_A	OCH	R_I	S_F	T_M	TAL
_AV	_MR	_TI	ANS	D_F	ER_	ISK	LIN	N_F	OCK	R_K	S_M	T_O	UND
_BE	_MO	_TR	AR_	D_M	ERA	KA_	LL_	N_M	OM_	R_M	SER	T_S	[illegible]
_BR	_O_	_UT	ARE	D_S	ERI	KAN	LLA	N_S	ON_	R_O	SKI	[illegible]	[illegible]
_EL	_OC	_VA	ARM	DA_	EST	KAR	LLE	NA_	ONG	R_P	[illegible]	[illegible]	[illegible]
_EN	_OD	_VI	AS_	DD_	ET_	KON	LAT	NDA	ONS	R_S	SOM	TAN	VAT
_F_	_PL	_VA	ASS	DE_	ETO	KT_	LAN	NDE	ONT	R_T	ST_	TEM	VEN
_FO	_PO	_X_	AST	DEL	FIL	L_F	M_F	NER	OR_	R_V	STA	TEN	VER
_FÖ	_PÅ	_YT	AT_	DER	FOR	L_M	M_M	NG_	ORM	RAD	STE	TER	VÄG
_I_	_RE	A_F	ATI	ED_	FÄR	L_S	M_S	NGA	PER	RAN	STO	TIL	X_M
_IN	_RO	A_S	ATT	EDD	FÖR	LAG	M_T	NGS	PLA	RE_	STR	TIO	YST
_KA	_SK	AD_	AV_	EL_	G_A	LAN	MAT	NIN	PRO	RED	STA	TNI	ÄGG
_KO	_SO	ADE	AVS	ELE	GAR	LAR	MED	NS_	PÅ_	RIN	STÅ	TON	ÄND
_LÅ	_SP	AL_	BEL	ELL	GD_	LAS	MEN	NST	R_A	RMA	T_A	TOR	ÄNG
_M_	_ST	ALL	BET	EN_	ILL	LAT	MIN	NT_	R_B	RME	T_F	TT_	ÖR_

Figure 2 KBS-CLASS tool creating a classification

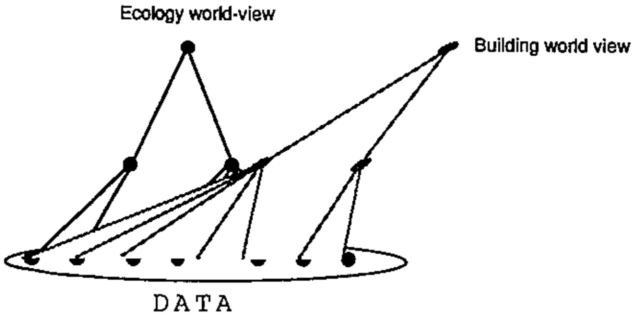


Figure 3 With a tool tailored to a familiar world view, the user can access data arranged according to a different view

Applications of the tool

Generally

The KBS-CLASS tool has several general applications.

1. To filter out information faster and more efficiently from familiar sources: filter building journals and electronic discussions.
2. To sort information already to hand: classify texts produced in-house and sort out documentation for buildings, e.g. building protocols.
3. To reach information previously beyond the horizon, that follows an unfamiliar system: (see Figure 3) find texts relevant to building in other sources than building sources, e.g. sources in chemistry and ecology and translate natural language search queries to classes for retrieval of pre-classified items such as CAD drawings, 3-D models and product descriptions.

Case study

Consider the following slightly futuristic scenario: you are required to determine the technical status of a building. You need this for making decisions on whether the building should stand and be refurbished or be demolished. Experts can be hired to analyse the building, but your organization wants a quick estimate of the costs for pursuing a number of different alternatives. If the building is to be demolished you are interested in knowing what compounds might be environmentally dangerous in the building. If it is to stand you are interested in reusing materials from other buildings. The documentation of the building is unstructured: it resides in a variety of different formats, written by different people.

The building was constructed some time ago and an effort has been made to scan texts and other documentation about the building. This information consists of papers produced before the construction of the building, during construction and in connection with repair and maintenance. There are free-text descriptions of what needs to be changed in your building, written down by a person who has left your organization and the region.

The information is structured in ways not appropriate to your needs. You need information on lifts, the materials used for joints, etc. and the information is presently classified as 'work protocols', 'materials lists' and 'report'. In order to make a quick environmental survey, you will consult a program available in the future called 'the environmental advisor'. This is a rule-based database that spots environmentally dangerous substances which can work with the KBS-CLASS tool and, indeed, with other tools.

You begin by putting the documentation you have into an unstructured database. With the aid of the KBS-CLASS tool you can now analyse the material. You need to find out what insulation has been used in the walls and you suspect there is asbestos in some building parts. Furthermore, you need detailed information on lifts and ventilation. Asbestos is mentioned in the report, but not its location. If the asbestos is safely hidden away from draughts and children, the regulations allow it to be kept there. If the building is to be demolished, the regulations require you to use safety precautions, concerning demolition and disposal.

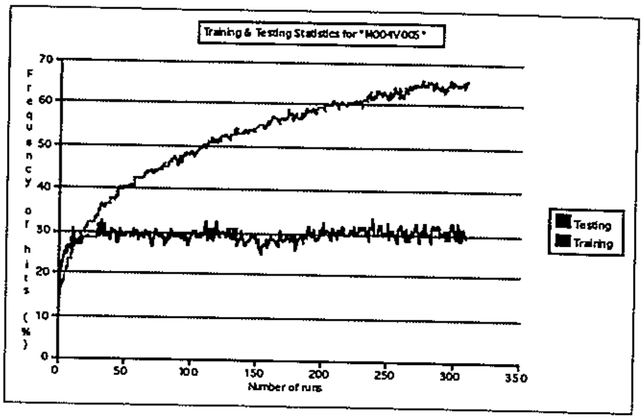
You use the environmental advisor to scan your database for dangerous materials. The environmental advisor can tell you what is dangerous in the things mentioned in your unstructured database, but not where they are in the building. The environmental advisor returns information on 45 dangerous materials. You select asbestos and ask the advisor to consult the KBS-CLASS tool. The environmental advisor sends information on those text parts in which it finds mention of the materials. The user interface of the KBS-CLASS tool then shows, with classes and illustrations (see Fig. 1), where the material is to be found. You ask to see the text. It contains a brand name that has been highlighted by the tool, based on information from the environmental advisor. Finally, you instruct the computer to send the information to your word processor, for completing the appendix section of your analysis and report.

Performance of the tool

The KBS-CLASS tool has been trained to distinguish between the 26 top classes of the BSAB System Product Table 1 (pointing at technical solutions and activity results).

Figure 4 shows the performance of the KBS-CLASS tool. The *y*-axis indicates the testing and training performance and the *x*-axis shows the number of training runs. One training run consists of all 8000 cases being presented to the network one at a time. Since the network is only nudged in the right direction after each case, several runs are needed. The testing performance, i.e. the ability of the network to classify texts it has never seen before correctly, quickly stabilizes at approximately 30% after a number of runs, as Fig. 4 also shows. After that training performance increases without improved testing. This apparent 'memory' effect could be explained by the network having too many hidden neurons. However, nets used with fewer neurons failed to train. Initially, the trigrams were the most frequently occurring in the texts. Testing was improved by almost 50% by adding a criterion that the trigrams selected should have a variance of 0.04. The performance of the network is promising if one bears in mind the short texts used and the numerous improvements that could be made. The texts used are terse product descriptions ranging in length from 80 to 240 characters. Longer texts would likely produce better results. Eight thousand product descriptions have been used as training cases for the tool. This is far too few given the network's size. As a rule of thumb, there should be the same number of training cases as there are connections in the network. This means that some 34 300 training cases would be needed. It is difficult enough to find 8000 cases, but this is likely to change as more building information becomes electronically based. Furthermore, the architecture of the network could be improved and so could the input representation. Two things will be considered for the future of the tool.

1. Improve the input: increase the number of training cases and improve the input representation (fewer and more relevant trigrams and pre-processing techniques).



settings:
training tolerance: 0.1
testing tolerance: 0.4
learning rate: 1.0

number of neurons:
input: 219
hidden: 140
output 26

Selection criteria for trigrams: Mean value of occurrence greater than 0.01 AND variance greater than 0.04

Figure 4 Training and testing perforance for the KBS-CLASS tool

2. Improve the network: adaptive learning (add neurons as needed), pruning (remove neurons not needed) and substitute ones and zeros for more efficient values (Fausett, 1994).

Future directions

An example of a possible work procedure has been presented in this article. However, this needs to extend to

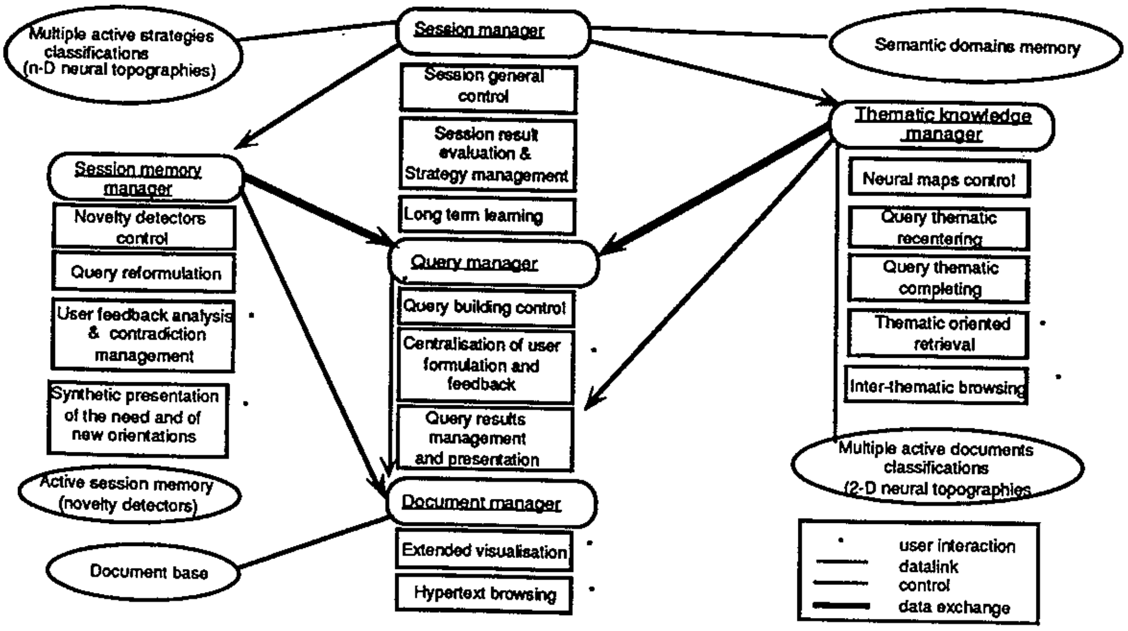

Multiple active strategies classifications (n-D neural topographies)
Session manager
Session general control
Session result evaluation & Strategy management
Long term learning
Semantic domains memory
Thematic knowledge manager
Neural maps control
Query thematic recentering
Query thematic completing
Thematic oriented retrieval
Inter-thematic browsing
Multiple active documents classifications (2-D neural topographies)
Session memory manager
Novelty detectors control
Query reformulation
User feedback analysis & contradiction management
Synthetic presentation of the need and of new orientations
Active session memory (novelty detectors)
Query manager
Query building control
Centralisation of user formulation and feedback
Query results management and presentation
Document manager
Extended visualisation
Hypertext browsing
Document base
user interaction
datalink
control
data exchange

Figure 5 Overview of the NOMAD system

more than just finding a suitable way of representing a classification system on the screen. The user is likely to want to mix and match different text retrieval methods, including Boolean operators (AND, OR, NOT, etc.) and relevance feedback and he or she will want to save search queries for later use. Some of these tasks could be performed with an improved design of the user interface while others must be handled by background processes. An example of an experimental design on a high level is the NOMAD system (Lamirel *et al.*, 1994) (see Fig. 5).

Conclusions

The KBS-CLASS tool points to a future in which an information system must be able to handle and integrate different knowledge representations at a high level. The tool has used a number of techniques: it has a *bottom-up* perspective that could be defined as 'knowledge from practice'. The knowledge of classification inside the tool is built up directly from the practice of classification used in the cases with which the tool has been trained. The tool uses a data representation in the form of word parts (trigrams) as input (Scholtes, 1993). This hopefully brings down the complexity of the input data and makes the tool less sensitive to noise, for example, misspellings. The author believes that the techniques are fruitful for further development. If one compares the KBS-CLASS tool with a prototype aircraft on a test flight, one could say that the tool manages to take off, but it does not fly very high. As soon as a tool with tolerable performance is available, attention will be moved to the integration of the tool in a larger context.

Acknowledgements

The project described in this paper was undertaken at the KBS-Media Lab at the Department of Structural Engineering, Lund University. KBS-Media Lab is a part of the Swedish National IT-building programme. The work was funded by NUTEK (the Swedish National Board for Industrial and Technical Development) within the IT-building programme.

References

Berglund, B., Christiansson, P., Hansson, B., Landin, A. and Modin J. (1992) *Kunskapsutveckling i byggprocessen.* (Swedish report on the Cube system). LUTVDG/(TVBP-3032), Lund University, Lund.

Cavazza, M. and Zweigenbaum, P. (1992) Extracting implicit information from free text technical reports, *Information Processing and Management*, **28(5)**, 609–18.

Christiansson, P. (1993) Dynamic knowledge nets in a changing building process, *Automation in Construction* **1(4)**, 307–22.

Christiansson, P. and Modin, J. (1993) Conceptual models for communicating knowledge in the building industry in *Proceedings from the MITC Conference*, Singapore, July.

Croft, W.B. (1992) Re: a national center for research in intelligent IR, *IR-L Digest Mailing List*, **IX(39)**.

Croft, W.B. (1993). Knowledge based and statistical approaches to text retrieval, *IEEE Expert*, **April**.

Fausett, L.V. (1994) *Fundamentals of Neural Networks, Architectures Algorithms and Applications*, Englewood Cliffs, NJ, Prentice Hall.

Harman, D.K. (ed) (1993) *The First Text Retrieval Conference (TREC-1)*. Gaithersburg, MD, NIST Computer Systems Laboratory, NIST Special Publication 500-207.

Jacobs, P.S. (1993) Using statistical methods to improve knowledge-based news categorization, *IEEE Expert*, **4/93, (8)** 13–23.

Kahle, B. (1991) *WAIStation User Guide, Prototype Version v 0.57.*

Lamirel, J.-C., Crehange, M. and Ducloy, J. (1994) NOMAD: a documentary database interrogation system using multiple neural topographies and novelty detection, In *Proceedings of the Third International ISKO Conference*, Copenhagen, 20–24 June, pp. 334–41.

Lawrence, J. (1991) *Introduction to Neural Networks.* Grass Valley, California Scientific Software.

Lelu, A. (1991) From data analysis to neural networks: new prospects for efficient browsing through databases, *Journal of Information Science*, 17(1), 1–12.

Rao, R. (1993). User research at PARC, electronic document, *PARC's World Wide Web-server.*

Rose, D.E. and Belew, R.K. (1991) A connectionist and symbolic hybrid for improving legal research, *International Journal of Man–Machine Studies*, **July**.

Schlieper, H. (1992). *Introduction to UN/Edifact Messages and Frameworks.* IBM Germany Information Systems GmbH, Germany.

Scholtes, J. (1993) *Neural networks, natural language processing, information retrieval*, doctorate thesis, University of Amsterdam.

Sillince, J.A.A. (1992) Argumentation-based indexing for information retrieval from learned articles, *Journal of Documentation*, **48(4)**, 387–405.

Tetzeli, R. (1994) Surviving information overload, *Fortune Magazine*, **11 July, 130**(1), 60–65.