

# Description of Events: An Analysis of Keywords and Indexical Names

Khurshid Ahmad, Paulo C F de Oliveira, Pensiri Manomaisupat, Matthew Casey and Tugba Taskaya

Department of Computing, University of Surrey  
Guildford, Surrey. GU2 7XH  
UK  
(k.ahmad@surrey.ac.uk)

## Abstract

Event modelling requires a good understanding of the modes used in communicating the events, including natural language, graphs and images. A case study of financial market movement, where text, or news wires, and graphical information, or a financial time series, were correlated, is described. This leads to a need for automatic text classification: a method based on unsupervised neural networks and autonomous assignment of keywords is described. These are preliminary results of an EU 5<sup>th</sup> Framework Project –GIDA (No. IST 2000-31123). Methods of corpus linguistics and terminology are used to underpin the methods.

## 1. Introduction

An event is defined as a significant occurrence or happening, or more specifically, as in physics, an event is a phenomenon or occurrence located at a single point in space time. In the late 20<sup>th</sup> century a tautological compound *news event* makes the meaning of the word ‘event’ even more explicit. A description of an event names the persons, places, things, or ideas in relation to the significant occurrence, happening or phenomenon: Osama bin Laden is frequently named in relation to terrorism; the Financial Times/Stock Exchange 100 (top companies) index (FTSE) is named in relation to the British, and possibly, EU economy; *relativity* and Einstein in relation to 20<sup>th</sup> century physics.

Reports of terrorism, stock market movements, and developments in theoretical physics, use written language, photographs, time series of financial transactions, graphs of key variables, and other symbol systems. Reports of events, political, economic, scientific or leisure, for instance, are crafted using a range of semiotic systems – from natural language to images, from time series to icons. An event, when described in natural language, involves the deliberate frequent use, and at times deliberate censoring of names related to the significant occurrence or phenomenon. For a specific event, described over a period of time in a number of texts, some persons, things or ideas are mentioned more or less frequently depending on their influence on the event. An event, perhaps at the lexical level of linguistic description, is a cluster of keywords or terms related to the specific area of human activity – terrorism, finance and commerce, physics, or football for example.

The names of (significant) persons, things or ideas act as an index to an event, an index which has linguistic rendering but can equally be referred through the use of other semes – images, graphs, mathematical symbols, circuit diagrams are some of the other indexical semes. Keywords-in-context (KWIC), largely common nouns sometimes qualified by adjectives, can be used to categorise documents related to a special subjects or, perhaps indirectly, to a specific events.

For us, event modelling requires an understanding of keywords and a collation of indexical names. For computer-based event modelling, involving information

extraction and retrieval, and text understanding, it is important (a) to automatically identify and verify new keywords and indexical names, (b) to be able to note nuances of, and changes in, use of the keywords and the indexical names, and (c) to correlate the information in text and in graphs through the use of indexical names and keywords.

News streams provided by organisations like Reuters or Bloomberg comprise a range of keywords and indexical names that may change from one news item to the next; an event modeller will need to filter the news from such a diverse information resource. Specialist information providers deliver not only news texts but also supply, for example, time series of changes in value of stocks, shares, currencies, bonds and other financial instruments.

We have a narrower focus than other authors in information extraction (see for example Gaizauskas et al, 1995 and Maybury et al, 1995) in that we are looking for changes in key financial instruments that are reported in financial news-wires. The news coverage of these instruments is of two types: first, there is a daily report about changes in the value (numerical) of the instruments for instance, one can see time series comprising historic data about the changes in values of currencies; second, the manner in which the value of the instruments changes depends on the reports relating, directly or indirectly, to the instrument. The reports, for example, about war or economic uplift/downturn, affect the value of the instruments. Some authors claim that there is a correlation between ‘good’ or ‘bad’ news relating to the instrument and its potential numerical value. In Section 2 we take this discussion further.

The news report is one of the most commonly occurring linguistic expressions. Despite being a good example of open-world data, a news report is a contrived artefact: each report has a potentially attention grabbing headline; the opening few sentences generally comprise a good summary of the contents of the report; there are *slots* for the date of origin and slots for photographs and other graphic material. This contrived artefact is highly focused and highly perishable, and usually contains references to one or more persons, places, events or actions. Automatic categorisation of news stories is of substantial interest to in a range of applications (Mani 1998) to information retrieval communities, and to major news vendors

supplying *on-line news*; Section 3 takes up this story further and we conclude in Section 4.

## 2. Keyword and Indexical Name Correlation

Generally, information is delivered to financial market operatives via electronic mail, newspaper, or company announcement briefings or company annual reports. Whatever its source, the information in the news is an important component in making investment decisions (Figure 1). Equally important are events like natural disasters or terrorist activities for example.

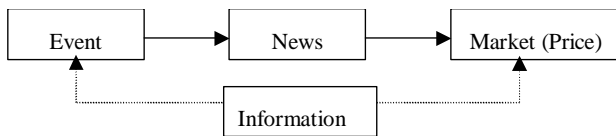


Figure 1: The relationship between Events, News and Markets (price) through Information.

For example, the terrible events of September 11, 2001, have had a catastrophic effect on financial markets world wide (See Figure 2). Various national economic indicators –indexical names – show the reaction on the date; there has been a decline in the value of these indices before that date and indeed a resurgence in the value afterwards.

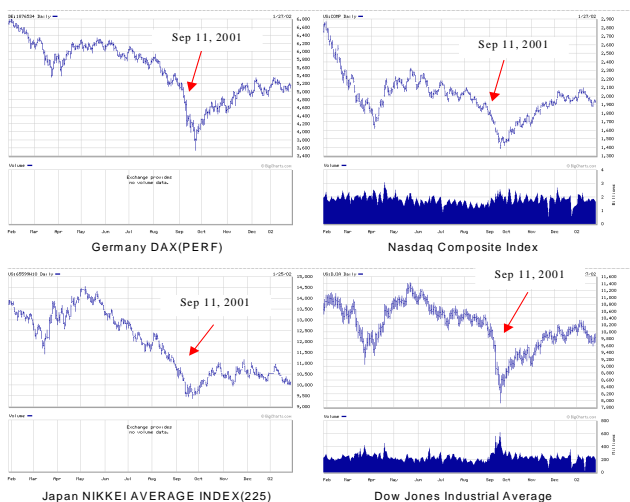


Figure 2: Movement from Feb 2002 to Jan 2002. Note the dip on and around Sep 11<sup>th</sup> 2001.

According to the Dow-Hamilton Theory (Rhea 1994), there are three kinds of price movements or market movements: (i) *Primary movement* which lasts from few months to many years and represents the broad trend within a market; (ii) *Secondary movement* last from a few weeks to few months and may sometimes be contrary to the primary movement; and, (iii) *Daily Fluctuations* can move with or against the primary trend and exist for a few hours to a few days.

### 2.1. Market Movement and Market Sentiment

Our work, sponsored in part by the EU-sponsored GIDA project (Project No. IST 2000-31123), focuses on primary movements. We report on some initial work that attempts to changes in an index, FTSE100, with changes in ‘market sentiment’ as expressed in news reports about the UK economy specifically and reports about the Wall Street indices. The later has substantial influence on the

UK economy. Financial analysts use sophisticated political, economic and psychological analysis to determine the reaction of market operatives and to predict the possible trading decisions of the operatives. Reports related to the sentiment use a range of metaphors to express the state of a market and its possible movements. Francis Knowles has written about the use of *health metaphors* used in the financial news reports: markets are full of *vigour* and are *strong* or the markets are *anaemic* or are *weak* (1996); most newspapers also use *animal metaphors* – there are *bull* markets and *bear* markets, the former refer to expansion, and indirectly to fertility, and the later to shy, retiring and grizzly behaviour much like that reported about bears in popular press and in literature for children. Indeed, there are fairly literal words that express the sentiment, as reported in the news wires, about the markets: financial instruments *rise*, *fall*, markets *boom*, *go bust*, and there are *gains*, *losses* within the markets, economies *slowdown*, suffer *slowturns*, whole industry sectors maybe *hardpressed*. Table 1 contains examples of good and bad news in a typical Reuters news stream:

Mainly Good News Stories	Rather Bad News Stories
Naval shipbuilder and military contractor Vosper Thornycroft has <b>boosted</b> its civil arm by buying facilities manager Merlin Communications (Nov 14, 2001)	Heavyweight banking and oil stocks have <b>dropped</b> up the leading share index as investors bet on fresh interest rate cuts.’ (Nov 21, 2001).
The FTSE 100 stock index looks set to open <b>stronger</b> today after Wall Street added to <b>gains</b> seen at the London close and with U.S. stock index futures boosted by rumours that Osama bin Laden had been captured.’(Nov 15, 2001).	The European Commission has <b>slashed</b> its official growth forecasts for the euro zone [...], predicting the most serious <b>slowdown</b> since the 1990s recession, with <b>lower growth</b> in 2002 than this year.’ (Nov 21, 2001).
Builder McCarthy & Stone has posted a 13 percent <b>rise</b> in annual pre-tax profits, built on <b>strong</b> sale prices for its retirement homes [...], but cautions that the <b>boom</b> may be over.(Nov 15 2001).	The FTSE 100 fell today, amid concern about how the U.S. economic <b>downturn</b> will hurt technology stocks and British Airways’ operations. (Dec 10, 2001).
Leading shares are expected to <b>rise</b> again after Wall Street steamed <b>higher</b> overnight and the market basked in a feel-good glow, dealers said.’ Nov 14, 2001).	Britain’s economy appears to be sailing along relatively smoothly despite the global <b>slowdown</b> and a string of high-profile job <b>layoffs</b> (Oct 22, 2001).
‘Leading shares have edged <b>higher</b> in early trade, <b>boosted</b> by <b>gains</b> in technology stocks in response to a Wall Street rally and <b>positive</b> expectations for the economic outlook.’ (Jan 4, 2002).	‘The <b>hard-pressed</b> manufacturing sector has recorded its biggest monthly production <b>drop</b> in almost a decade, <b>sinking deeper</b> into <b>recession</b> . (Nov 5, 2001).

Table 1. Examples of ‘good’ and ‘bad’ news stories in Reuters News Wires (Oct 2001-January 2002)

The above table contains examples of how the market is moving. But here we have free natural language complete with ambiguity and nuances of meaning: so there maybe a ‘rise in profits’ and a ‘strong sale prices’, in the story about builder’s McCarthy & Stone above, both phrases suggesting that this is a good news story, except for the last sentence suggesting that ‘boom maybe over’. Nevertheless, many of the news items do not change the nuance of the story by such highly temperate notes.

## 2.2. Correlating Sentiment and Market Indices

We created a corpus of 1,539 English financial texts from one source (Reuters) on the World Wide Web, published during a 3 month period (Oct 2001-January 2002) comprising over 310,000 tokens. The corpus comprised a blend of both short news stories and financial reports. Most of the news is business news from Britain with thirty percent of the news is from Europe and from the United States.

We found over 70 terms each for conveying good news and bad news in the above corpus. The texts in our corpus were also time stamped, and by using our text and terminology management system, System Quirk, we computed the cumulative weekly frequency of *good* words and *bad* words during one month – November 2001. The ‘week’ is a working week comprising 5 days, Mondays-Fridays:

Time (5 day Week)	Good Word Frequency	Bad Word Frequency
1	<u>58</u>	40
2	71	<u>75</u>
3	77	66
4	73	59
5	72	<u>28</u>
<b>Total</b>	<b>351</b>	<b>268</b>

Table 2: Frequency of Good and Bad words in Nov 2001. The underlined figures in the 2<sup>nd</sup> and 3<sup>rd</sup> columns indicate the minimum value of the frequency and the numbers in italics are the maximum value.

Table 2 shows that in November the highest frequency of ‘good’ words was in week 3 (77 instances) and the ‘bad’ words was in week 2 (75 instances). How does this correlate with the movements of the London stock and shares as expressed by the FTSE 100? Figure 3 provides an example of the correlation between the frequency of ‘good’ words from news in November in our corpus and close prices of FTSE100 Index for the whole month of November.

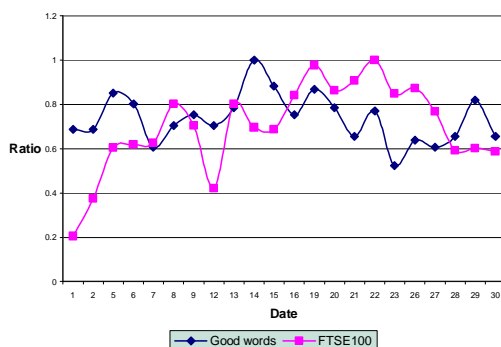


Figure 3: Market correlation between ‘good’ word frequency and FTSE index.

The highest value of the FTSE 100 index was on on 22 November 2001 (5345.94). There is a perhaps a correlation between the changes in the value of the index and the frequency of ‘good’ words: Positive gradient in

the ‘good’ words time series correlates well with the positive gradient in the FTSE 100 index values. What will be interesting for the purposes of predicting the movement of the market, will be a correlation that suggests that a rise in the number of good words one day nudges the market. Correspondingly, that a decrease in the number of the previous day will lead either to a static market or falling market the next day. The same can be said, perhaps in reverse, about the bad news words (see Figure 4).

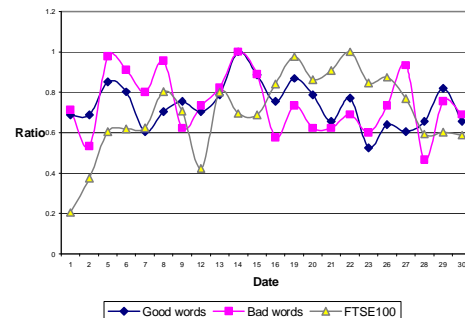


Figure 4: Good and bad word frequency correlated with FTSE 100.

Figure 4 shows ‘good’ word and ‘bad’ word frequency is perhaps correlated with FTSE100 values. For example, from 23<sup>rd</sup> to 29<sup>th</sup> November the frequency of ‘bad’ words increased while the FTSE100 went down over this period. After the 29<sup>th</sup> November, the FTSE100 value slightly increased as the ‘good’ word frequency also increased.

The above analysis and the concomitant results are of a tentative nature in that work is progressing in three major directions. First, one needs a bigger corpus, and a longer time series, to be more assertive about a correlation between an index and the corresponding sentiment-bearing terms. Second, further analysis is underway to note that the good news is sometimes tempered with bad news and vice versa – this will involve a phrasal or sentential analysis. Third, the notion of a ‘time series’ is a carefully defined concept for a series of cardinal numbers collected at discrete intervals of time or collected continuously; we are exploring the status of a time series made up of counts of lexical strings found in a news report that may have been produced over an approximate time. Nevertheless, Figures 3 and 4 show how a news stream, comprising subject specific texts, can be visualised especially in the context other indexical names.

## 3. Classifying News Wires Using Keywords

### 3.1. Categories of News Reports

A news stream comprises news stories that: (a) range over a whole range of subjects; (b) the news may emanate from or maybe about a nation state; and (c) the news may be focused on a certain specific area of human enterprise Reuters labels for items (a)-(c) are ‘Topic’, ‘Country’ and ‘Industry’; these labels are used by Reuters’ sub-editors to tag each news story with one or more Topic and Country tags, and in some cases with the Industry tags. These pre-assigned tags, about 1000 different tags in all, can, in principle, be used to categorise individual news stories in a news stream. However, the plurality of tags, that is the presence of one or more tags with either Topic, Country,

or Industry, makes such a categorisation more complex. Before we discuss how to deal with such a complex categorization task (see Section 3.5), which is possibly subjective in that the categories are based on an ontology which was created by Reuters themselves, we look at how to categorize texts based on (semi-)automatically extracted keywords.

One well-recognized way of describing news reports is to classify the texts as a distinct *register* or *genre* of writing. The term register is used to indicate that the language within a specialized field differs from that of *general language* or language of everyday use, at lexical, syntactic and semantic levels. A large collection of general language text may thus be contrasted with a set of specialist reports at various linguistic levels, including lexical and semantic.

An important use of this contrast is in a method of semi-automatically identifying the terms of a set of specialist domains. This method involves comparing the frequency of systematic terms in a collection of specialist texts sometimes called a *corpus*, with the frequency (or absence) of the terms in a carefully compiled corpus of general language texts. Each term can be construed as a dimension in a vector space and the presence or absence of a term within a text is then used to allocate the text its position within the vector space. There is some evidence from work in linguistics that word categories (nouns, verbs, adjectives, adverbs, prepositions, etc.) may be inferred from the statistical occurrences of words in different contexts. For Kohonen and his colleagues, “*context patterns*” consist of groups of contiguous symbols; the authors cite pairs or triplets of words in a sentence as an example of such patterns. Such *pairs* or *triplets* are then used as inputs in the training and testing of a neural network (the so-called self-organising feature maps or SOFM; details of this map is presented in the next Section 3.2). Kohonen has shown that a SOFM-trained word context pairs, derived from 10,000 random sentences, shows ‘a meaningful geometric order of the various word categories’. A larger SOFM, the WEBSOM has been variously described by Kohonen as a *scheme*, *content-addressable memory*, *method* and *architecture*. WEBSOM is a two-level self-organising feature map comprising a word category map and a document category map, which has been used to classify newsgroup discussions, full-text data and articles in scientific journals (Kohonen 1997b, Kaski *et al.* 1996). Terms were pre-selected by the builders of WEBSOM. There are other neural network architectures that have been used in text categorisation, especially the widely-used supervised learning algorithms – SOFM is based on unsupervised learning algorithm – which have been discussed by Lewis (1995).

Consider a set of texts that may have been selected according to certain criteria: for instance, all texts streaming along a news wire over a short period of time comprising news related to specialist topics – like environmental news or economic news. Such a short news stream may contain may result in a text collection, or if collected systematically, a text corpus, that may be characterised the high frequency of environment – or economics – related terms. However, over a long period of time this may not be the case as the news stream may start to deliver texts in different specialist areas. So how do we extract terms from such a corpus?

Specialist texts can be distinguished from a general language text at the lexical level of linguistic descriptions by looking at the ratio of relative frequency of a linguistic token in a specialist text and its frequency in general language texts. This ratio has been termed *weirdness* to indicate how it measures the preponderance of words in specialist texts that would be unusual in general language, (see, for example, Ahmad 1995).

Typically, before text documents are represented as vectors in order to act as the input to a text categorisation system, pre-processing takes the form of filters to remove words ‘low in content’ from the text (see the WEBSOM method in Kaski *et al.* 1996). We remove punctuation, numerical expressions and *closed-class words* as a precursor of generating the feature set. Vectors representing news texts were created on the basis of a lexical profile of the training set of texts. This lexical profile was determined by two measures: the frequency of a term; and, a weirdness coefficient describing the subject-specificity of a term.

The feature set was created by first selecting the top 5% most frequently occurring words, and from this set, by choosing the words with the highest weirdness coefficient. Subsequently, the 50 most frequent words are selected, excluding spelling mistakes, and numerical expressions and terms too infrequent to provide consistency within a domain are avoided. A high value for the weirdness coefficient is indicative of a word which is uncommon in general language but common in the specialist corpus under examination and is thus a good candidate for a domain term or other word specific to that genre. By disregarding words with a weirdness coefficient lower than a threshold, many *closed-class words* and other terms common in general language are automatically removed. Before we show texts can be categorised using the above method, we digress to briefly outline the Kohonen Self-organising Maps

### 3.2. Kohonen Self-organising Maps

A SOFM is a neural network and associated learning algorithm that is designed to produce a statistical approximation of the input space by mapping an input in to a two-dimensional output layer (see Kohonen 1997a for an extensive discussion). The approximation is achieved by selection of features that characterise the data, which are output in a topologically ordered map. The Kohonen Self-Organising Map has a close resonance with the *k-means clustering* method, with the additional constraint that cluster centres are located on a regular grid (or some other topographic structure). Furthermore their location on the grid is monotonically related to the pair-wise proximity (Murtagh & Hernández-Pajares, 1995).

The basic SOFM consists of a single layer of neurons formed into a two-dimensional lattice. Each neuron is connected to the input via a set of connections utilising connection weights, just as in a perceptron. There is no ‘output’ of the map, rather the values of each neuron’s weight vector are used to visualise the formed topological ordering. The weight vectors form a cluster prototype that is measured against each input to determine how ‘close’ the vector is to a given cluster. Since the map is two-dimensional and the input typically has a high dimensionality, the SOFM acts as a dimensional squash

allowing the visualisation of features within multi-dimensional data.

Learning is achieved in the SOFM using a competitive algorithm. The Euclidean distance between each training input vector and all weight vectors is determined. The neuron with the weight vector that has the smallest Euclidean distance to the input pattern is termed the winner. To reward the winning neuron its weight vector is adjusted to be ‘closer’ to the input vector, with the amount of adjustment determined by the number of times the training patterns have been presented (via the learning rate). Additionally, all vectors within a defined neighbourhood of the winner are adjusted, essentially forming a cluster of similar values that are seen to be activated by the winner. The neighbourhood size decreases with the number of training cycles, typically using a bubble neighbourhood (a rectangular area) or a Gaussian neighbourhood, both centred on the winning neuron. The adjustment of the weight vector towards the input is achieved by effectively ‘moving’ the weight vector’s direction towards that of the input. This simple process of adjusting ever-smaller neighbourhoods of winners allows the formation of clusters within the lattice. As the number of cycles increases, the clusters become more stable and can be viewed through probing to find winners using test data.

The principal way in which information about the clustering performed by the SOFM learning algorithm is visualised is through probing with a test set to find the winning neurons. The co-location of different winners from different categories highlights the similarity between clusters. The effectiveness of such clusters can be measured by comparing different versions of the map trained on the same data through a technique being developed by Ahmad et al (2001), where Fisher’s Linear Discriminant Rule is used to quantify the discrimination ability of different clusters.

### 3.3. Limitations of a SOFM

The SOFMs strength lies in its ability to *statistically* summarise the input space. However, it has been shown that the basic SOFM does not always produce a *faithful* approximation (Ritter & Schulen, 1986). This faithful approximation is defined as the proportionality between the density of the weight vectors and the density of the input space. Lin et al (1997) has shown that the SOFM underrepresents high-density regions and overrepresents low-density regions.

### 3.4. Automatic Categorization of Texts Based on Keywords Using an SOFM

Our text corpus consisted of 100 Associated Press (AP) news wires selected from 10 pre-classified news categories shown in Table 3 together with their icons. The average length of the articles was 622 words.

### Text Categories











1	Bioconversion		6	Exportation of Industry	
2	Pollution Recovery		7	Foreign Trade	
3	Alternative Fuels		8	Int. Drug Enforcement	
4	Fossil Fuels		9	Foreign Car Makers	
5	Rain Forests		10	Worldwide Tax Sources	

Table 3: Text categories used in the TIPSTER – SUMMARY program

The 100 AP news wires comprised over 56,000 words. System Quirk was used to compute frequency distribution of words in the AP News wire corpus. The System also has access to the frequency distribution of words in the British National Corpus (Aston and Burnard 1998) a carefully compiled general language corpus. Some of the high weirdness terms, e.g., *drug*, *taxes*, *pollution* and *environmental* are important keywords, but the same cannot be said for ‘terms’ like *billion*, *percent* and *federal*. Usually, proper nouns are also flagged as terms by this method. The feature words identified for the 100 AP News Wire texts are shown in Table 4 according to rank:

1	percent	15	congress	28	dioxide	41	corp
2	tax	16	mexico	29	marine	42	forests
3	billion	17	emissions	30	mazda	43	cocaine
4	drug	18	drugs	31	gases	44	enforcement
5	reagan	19	fuels	32	shale	45	warming
6	cars	20	senate	33	deficit	46	smog
7	taxes	21	auto	34	export	47	ozone
8	environmental	22	proposal	35	recycling	48	Massachusetts
9	pollution	23	gasoline	36	epa	49	imports
10	fuel	24	exports	37	honda	50	automobile
12	federal	25	vehicles	38	methanol	51	trafficking
13	dukakis	26	ohio	39	automakers		
14	bush	27	green-house	40	panama		

Table 4: Feature words identified for the 100 AP News Wire Texts.

Having identified the feature set the training vectors for each of the texts could then be generated. Each vector consisted of binary values indicating the presence or not of each of the feature words determined above.

We have developed a system for creating Kohonen Feature Maps (SANC: Surrey Artificial Network Classifier). The system, after having trained an SOFM, is also capable of testing it. (There are facilities to vary the key parameters associated with the learning algorithm).

The system can be used to test the trained. Furthermore, the system allows the storage of previously trained maps for reference purposes (Ahmad, Vrusias and Ledford 2001).

The results of the Kohonen classifications for full texts are shown in Figure 5. Using symbols to represent each of the locations of the ‘winning node’, the position of each text is indicated across the two-dimensional map (shown in Table 3). It can be seen that the quality of clustering for the full-texts is successful for a range of categories, but especially for categories 9 (FOREIGN CAR MAKERS) and 10

(WORLDWIDE TAX SOURCES). Patterns in categories 1 (BIOCONVERSION), 4 (FOSSIL FUELS), 6 (EXPORTATION OF INDUSTRY) and 8 (INTERNATIONAL DRUG ENFORCEMENT) are also effectively grouped together. The widespread distribution of Class 5 (RAIN FORESTS) shows it to be the worst class on the map.

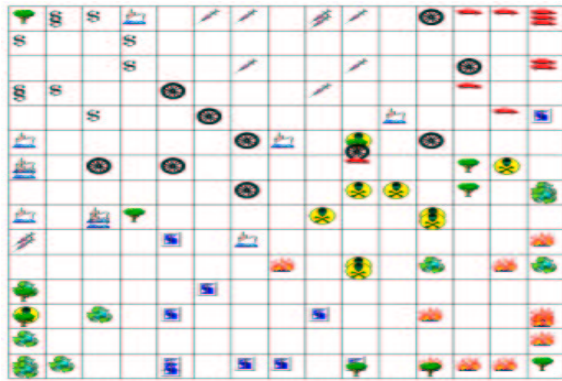


Figure 5: Results of a Full Text Map trained using exponentially decreased neighbourhood and learning rate.

These results for the trained Kohonen map were similar across a number of trials despite variations in training method and learning rate used. Some categories, for example 10 (WORLDWIDE TAX SOURCES), clustered consistently better than others for instance 5 (RAIN FORESTS). By simply counting the number of feature set words that appear in at least nine of the ten texts of each category, the best clustered categories are guaranteed to have some of these words. This reflects the tendency of these categories to cluster well. On the other hand, for a category in the 'best' case, only four of the texts share a common feature set word. This difference in classification difficulty was also seen in the TIPSTER results from two human assessors.

### 3.5. Multiple Categories and Text Categorization

Recall that Reuters News Agency has three categories: "Topic", "Country" and "Industry". The total number of different tags, or concepts, defined in these three categories is approximately 1000.

We have created a text corpus of 800 news stories streamed by Reuters in 1997. Each of the news stories is encoded in XML format and has clearly delineated headline, date, writer, text and code fields using XML tagset. The XML-based delineation helps in extracting keywords associated with the Topic, Country and Industry tags. The frequency of each concept was calculated within 800 documents; 80 of the keywords turned out to be more frequent than other 920: the distribution of the keywords in the various fields was as follows:

<b>Industry</b>	<b>39</b>	<b>Topic</b>	<b>32</b>	<b>Country</b>	<b>19</b>
-----------------	-----------	--------------	-----------	----------------	-----------

A SOFM was trained for categorising the 102 out of the 800 news stories. The input vector was created from the 80 most frequent keywords associated with the triple, Industry-Topic-Country: the absence and presence of a particular keyword was used to create the input vector for each of the texts. The neural network was trained 100 times. The vector thus created can, in principle, cope with

upto 39 different categories of 'Industry', of 32 different 'Topics' and '19' different countries. The downside here is that documents comprising references to the 920 keywords may not get classified as well as those that may comprise the 80 categories used in the construction of the input vector.

After the training period, the pre-specified Reuters documents were visualised on the map. As can be seen in Figure 6, the documents associated with each neuron were represented by a blue square. The distribution and the similarity of the documents were based mostly on the "Topic". On the lower right side of the map, the topics related to "Government/Social" were clustered. The subtopics of "Government/Social", for example "Sports" and "Art", were also clustered near this area. The documents categorised as "Management" were found on the lower left corner of the map. "Strategy and Plans", "Comments/Forecast" and "Economy" follow this as we approach the upper left corner. "European Community" documents were found on the upper right corner of the map.

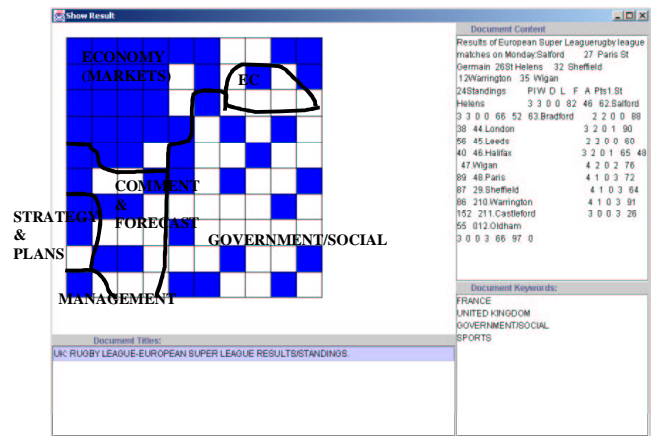


Figure 6: A Categorisation of Reuters news stories using pre-specified category information.

## 4. Afterword

Our current work involves evaluating the categorisation produced by the method that relies on different distribution of specialist terms in special and general language texts with that of using networks to classify texts that have pre-specified category information as was the case just described.

The pre-specified categories appear to be complex and, as mentioned above subjective in nature. We are currently examining whether a summary of text may give us some indication of the category. The reasoning is as follows: a full news story may contain extraneous material and a good summary will eliminate sentences within the text that are not directly related to the category or categories. Lexical cohesion studies have shown that keywords form the glue that helps to create a cohesive and coherent texts (Hoey 1991). In our previous work on AP news wires (Ahmad, Vrusias and Ledford 2001) we looked at three different types of text streams – headlines only, news summaries and full news items and categorised these texts using self-organising feature maps (SOFM). We found that an SOFM trained on vectors related to summaries

only provides a fairly accurate cluster when compared with vectors related to full text. This work is currently being carried out on the 102 Reuters texts mentioned above.

An analysis shows a vector for the 102 texts using our method based on the weirdness of the keywords within the news stories (Table 5).

Element	Description	Words
1 – 25	<i>Single Words</i> Top 25 simple words with high weirdness and high frequency	inventories, yen analysts, merger cents, peso investors, exports quarterly, forecast pesos, shares dealers analyst directive billion, soccer traded, allegations trading, stocks fiscal, tobacco nickel, earnings
26 – 30	<i>Compound Words:</i> 5 most frequent compound words	shareholder newsroom worldwide online chairman
31 – 40	<i>Proper Nouns</i> 10 proper nouns with high weirdness and high frequency	dorfman compuserve novell aol microsoft ec kimberly chrysler saudi netherlands
41 – 45	<i>Movement Indicators:</i> 5 most frequent downtrend words.	lost fall falling risk losses
46 – 50	<i>Movement Indicators:</i> 5 most frequent up trend words.	up growth high added strong

Table 5: Vector for the 102 Reuters news items (c.1997)

Note that in the above vector we have included movement indicators, proper nouns and compound words together with the single word terms. The 30 keywords and 10 proper nouns/indexical terms, together with 10 movement indicators will help us to define an event. Initial results of this analysis are encouraging in that we obtain the major clusters much like as found in Figure 6

We are currently exploring the notion that news streams will be filtered by using a trained Kohonen SOFM and the filtered text will be used to study market movement. The filter has to be 'cleaned' in that news stories are perishable items with constantly changing subjects – one idea is to re-train the network everyday, towards the end of the day perhaps, with a fixed number of stories which will exclude the very first day of the previous training set and include yesterday's news stories.

Event modelling, especially in noisy and dynamic environments, requires a careful consideration of the key concepts, expressed as keywords, and of indexicals like persons, places, things or ideas which play a crucial role in turning an occurrence, happening or phenomenon into a significant one.

## References

- Ahmad, K., Vrusias, L. & Ledford, A. (2001). Choosing Feature Sets for Training and Testing Self-organising Maps: A Case Study. *Neural Computing and Applications*, 10(1), 56-66.
- Ahmad, K. Pragmatics of Specialist Terms and Terminology Management. (1995). In (Ed.) Petra Steffens. *Machine Translation and the Lexicon*. pp. 51-76. Heidelberg: Springer.
- Aston, G. and Burnard, L. *The BNC Handbook: Exploring the British National Corpus with SARA*. 1998. Edinburgh: Edinburgh University Press.
- Gaizauskas, R., Wakao, T., Humphreys, K., Cunningham, H. & Wilks, Y. (1995). Description of the LaSIE system as used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. Morgan Kaufmann.
- Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford: Oxford University Press.
- Kaski, S, Honkela, T, Lagus, K & Kohonen T. (1996). Creating an order in digital libraries with self-organising maps. In *Proc. WCNN'96, World Congress on Neural Networks, 1996*, pp 814-817. Lawrence Erlbaum and INNS Press.
- Knowles, F. (1996) *Lexicographical Aspects of Health Metaphors in Financial Texts*. In (Eds.) Martin Gellerstam et al. *Euralex'96 Proceedings (Part II)*. Göteborg, Sweden: Göteborg University. pp 789-796.
- Kohonen, T. (1997a). Exploration of very large databases by self-organizing maps. In *Proceedings of ICNN'97, 1997*, pp. PL1-PL6, IEEE Service Center, Piscataway, NJ.
- Kohonen, T. (1997b). *Self-Organizing Maps*. 2<sup>nd</sup> Ed. Berlin, Heidelberg, New York: Springer-Verlag.
- Lewis, DD. (1995). Evaluating and optimising autonomous text classification systems. In *SIGIR 95: Proc. of the 18<sup>th</sup> Annual ACM-SIGIR Conference on Research and Developments in Information Retrieval*. pp 246-254.
- Lin, J.K., Grier, D.G. & Cowan, J.D. (1997). Faithful Representation of Separable Distributions. *Neural Computation*, 9(6), 1305-1320.
- Mani, I. (1998) *The TIPSTER SUMMAC Text Summarization Evaluation*. Mitre Technical Report: MTR 98W0000138, 1998.
- Maybury (1995). Generating Summaries from Event Data. *Information Processing and Management*. 31(5), 733-751.
- Murtagh F., Hernández-Pajares M. (1995). The Kohonen Self-Organizing Map Method: An Assessment. *Journal of Classification*, 12, 165-190.
- Rhea, R. (1994). *The Dow Theory*. Burlington: Fraser Publishing Company.
- Ritter, H. & Schulten, K. (1986). On the Stationary State of Kohonen's Self-Organizing Sensory Mapping. *Biological Cybernetics*, 54, 99-106.