

# A Text Mining Approach to Tracking Elements of Decision Making: a pilot study

C. Chibelushi, B. Sharp, A. Salter

School of Computing, Staffordshire University, Beaconside, Stafford ST18 0AD

Email: c.chibelushi@staffs.ac.uk; b.sharp@staffs.ac.uk; a.m.salter@staffs.ac.uk

**Abstract.** Understanding rework, the causes of rework, and the relationship between issues, decisions and the associated actions, is crucial in minimizing the fundamental industrial problems in system engineering projects. The aim of our research is to apply text mining techniques to track elements of decision making and extract semantic associations between decisions, actions and rework. Text mining is similar to data mining: while data mining seeks to discover meaningful patterns implicitly present in data, text mining aims to extract useful information and discovering semantic information hidden in texts. This paper describes work carried out as part of the first phase of our research, and investigates the effectiveness of using the text mining technique of lexical chains to identify important topics in transcribed texts. This research is part of the TRACKER research project which studies rework in system engineering projects.

**Keywords:** text mining, lexical chaining, tracking issues, rework.

## 1. Introduction

In the face of shrinking budgets, software engineers can no longer afford costly rework. By understanding rework and its causes, and tracking the issues, decisions and the associated actions we can help minimise the fundamental industrial problem of rework in system engineering projects. Rework occurs in situations where a particular decision re-addresses issues or actions resulting from a previous decision. Revisiting any of these decision making elements can be for positive or negative reasons<sup>1</sup>. Another factor influencing rework is the nature of note taking in meetings. Whittaker *et al.* (1994) and Minneman & Harrison (1993) found that meeting recorders did not have enough time to take adequate notes, and as a result they were not only restricted from participating in the meeting by their note taking task, but also wrote notes that were too terse, thereby missing important details. This has often led to unnecessary repetition, inappropriate or incorrect recording of decisions, and the use of a significant amount of time in meetings. We also believe rework through changing or unplanned software requirements is inevitable in large projects. However, a significant amount of rework arises as a result of communication failures between decision makers and participants during meetings (Rayson 2003).

To date, market and research emphasis has been placed on multimedia, video conferencing, decision support systems, and computer-supported cooperative work (Kazman *et al.* 1995). These systems are normally used in discussions, information analysis, and as a means of reference, before a particular decision is made. What is required is a method which assists decision makers in detecting, tracking and organising the enormous amount of data obtained from these systems so as to minimise or reduce rework.

Decision making can involve the analysis of large volumes of textual information including that relating to previous decisions and actions, and the undertaking of a series of consultations. We believe that most decisions are made through meetings and so it is important to utilise the information contained in both the audio recordings from the meetings (herewith referred to as transcripts) as well as the resulting minutes of the meeting. In our research both transcripts and minutes of the meetings are analysed with the view to

---

<sup>1</sup> A **negative reason** could be when the decision makers have no knowledge of similar decisions being made previously (due to different reasons such as losing the meeting documents, or the key decision maker leaving the job, etc.) in this case, the decision makers must decide on the same thing again. A **positive reason** could be when the decision makers re-visit a particular decisions to make a new decision that will improve the current situation.

identifying the key issues and actions discussed at these meetings, and discover the semantic relationships between decision, issues, actions and rework.

The research described here is part of the TRACKER project which aims to produce a semantic model that can help to reduce rework and increase efficiency in decision making within system engineering projects. We are also developing a set of tools (as a recall aid) to give better support to managers and participants during a meeting by providing a historical development of their decisions and associated actions. In this paper we focus on the analysis of transcripts recorded in meetings to extract key elements of the decision making process, namely issues, decisions, and actions initiated by the participants at the meetings. To this end we are applying text mining techniques to infer key issues and identify the set of actions associated with each of the decisions discussed at the meeting.

The paper is organised as follows. Section 2 describes an adaptation of the CRISP-DM model used in text mining while section 3 reviews related work to text mining. Section 4 describes our approach to the use of text mining in the tracking of issues. Section 5 provides some experimental results from the application of the approaches outlined in section 4 and section 6 highlights some of the problems in the application of the approaches and suggests future work.

## 2. Text Mining using the CRISP-DM Model

Text mining is an emerging research area concerned with the process of extracting interesting and non-trivial patterns or knowledge from text documents (Tan 1999). The aim of text mining is similar to data mining in that it attempts to analyse texts to discover interesting patterns such as clusters, associations, deviations, similarities, and differences in sets of text (Hidalgo *et al.* 2002). Liddy (2000) considers text mining to be a “sub-speciality of the broader domain of Knowledge Discovery from Data (KDD), which in turn can be defined as the computational process of extracting useful information from massive amounts of digital data by mapping low-level data into richer, more abstract forms and by detecting meaningful patterns implicitly present in the data”.

Our text mining approach combines both data mining and natural language processing methods to analyse transcripts in order to discover patterns that identify the elements of decision making, namely issues, actions, the initiators of decisions, and any hidden connections between initiators, decisions and rework. We also based our text mining approach on the Cross-Industry Standard Process for Data Mining (CRISP-DM), which is a comprehensive methodology, developed in 1966 and adopted by many researchers in data mining. Although CRISP-DM is specifically designed for data mining projects (Chapman *et al.* 2000), we have adapted the phases of the CRISP-DM, illustrated in Figure 1, to our text mining approach which is shown in Figure 2.

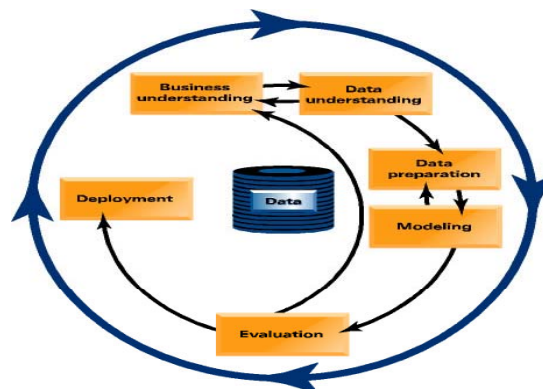
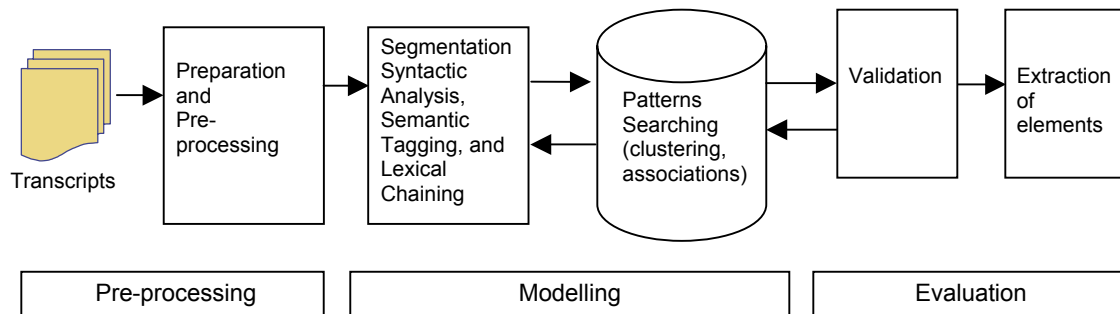


Figure 1. Phases of the CRISP-DM reference model (Chapman *et al.* 2000)

We identify three main phases: a pre-processing phase, a modelling phase and an evaluation phase. The pre-processing phase is an important phase and common to both data and text mining; it consists of identifying the business objectives and preparing the data for processing. Here we start by identifying the data (which consists of transcripts and the minutes of meetings), and clarifying the aim of the analysis. After exploration and verification of the transcripts’ quality we then prepare the transcripts for linguistic

analysis. This involves the removal of redundant and illegal characters, ambiguous characters, incorrect hyphenation, and the resolution of ambiguous punctuation. In the modelling phase we apply various techniques mostly derived from natural language processing such as text segmentation, semantic tagging and topic tracking. In this phase we do not aim at developing a model of the text as advocated by the CRISP-DM model, instead we apply a series of classification techniques to discover the linguistic patterns and semantic conceptual constituents used by participants in order to extract key elements of the decision making process, namely issues, decisions and associated actions. We also identify initiators of these decisions and actions. This phase is discussed further in the next section. In the evaluation phase we validate the syntactic patterns and the semantic associations that describe the elements of decision making and rework.



**Figure 2.** Text Mining Phases

As this work is part of a larger project, the extracted elements of the decision making process are then processed by a tracking tool which records these elements and provides search and navigation facilities to assist users in linking issues, decisions and actions. By providing a semantic link to previous decisions, issues and actions to the participants we make the decision making process more transparent, and if greater transparency can be achieved then the number of examples of incorrect or inefficient rework could be reduced (Salter *et al.* 2004).

### 3. Related work in text mining

Current methods in text mining use keywords, term weighting, or association rules to represent text content. Keywords have been used to explore connectivity and consistency within a collection of documents. Sardinha (1995) and Ellman & Tait (2000) used keywords to explore the connectivity and consistency within texts. Sardinha (1995) used keywords in the testimony of major witnesses in the OJ Simpson trial. Natural language is ambiguous and the same keyword may express entirely different meanings, e.g. “Washington” may represent a person or a place. The semantic disambiguation of such polysemous words is normally resolved through context. The inverse problem is that different expressions may refer to the same meaning, e.g. “car” and “automobile”. From these two problems, it is easy to rule out the surface expression of the keywords alone as a proper representation for text mining for our application.

When using the term weighting technique, document representation is done by first removing functional words (e.g. conjunctions, prepositions, pronouns, adverbs, etc.) and then assigning weights to content words (e.g. agent, decision making), in an attempt to describe how important the word is for that particular document or document collection. This is because some words carry more meaning than others. In addition to a basic division of the text into function and content words, some systems like ARROWSMITH (Weeber *et al.* 2001) go beyond a single word unit for the content words as they are more meaningful, for example ‘oxygen deficiency’ carries more meaning than the single words, ‘oxygen’ or ‘deficiency’ on their own.

We also argue that important information is conveyed with functional words, e.g. negations, and adverbs. Such words are usually treated as stop words, so by removing them from the analysis the relationships and cohesiveness between words are lost. Moreover, the current term weighting technique does not discriminate between information conveyed with higher level sentential structures, such as the interrogative and imperative information. Such information is crucial in recognising communicative intentions as they capture special communicative acts such as questions, requests, complaints and recommendations, which are especially useful when exploring decision making activities.

Association rules are popular representations in data mining but have also been used in text mining. An association rule is a simple probabilistic statement about the co-occurrence of certain events in a database or large collection of texts. For example, FACT is a system developed by Feldman & Hirsh (1996) which finds associations or patterns of co-occurrence amongst keywords describing the items in a collection of texts. Unlike numerical data processing, transcripts of meetings and minutes of meetings require overcoming considerable difficulties due to the unstructured nature of the textual information. Consequently, such an approach would necessitate both a quantitative and semantic analysis of the transcripts.

## 4. Our approach

In this paper we shall describe the work undertaken in analysing the transcripts of the meetings. Having defined the objectives of our text mining, we recorded a set of meetings and transcribed them for further processing. These transcripts were manually edited to prepare for the modelling phase. In this phase, we are primarily concerned with analysing the content of the transcripts in order to track the themes discussed and extract the key issues and any associated actions as well as identifying the initiator. To this end our approach combines statistical natural language processing and semantic analysis of the transcripts.

Our study utilised five transcripts ranging between 8,314 and 20,900 words. For the purpose of this paper, we shall focus on one transcript containing 8314 words (referred to as transcript 'A'). This transcript involves four meeting participants discussing how to develop two text analyser modules: an XML converter, and a Systemic Functional Linguistic Lexico-grammatical analyser.

In the following sections we describe the modelling phase in our text mining approach, which involves three major tasks: syntactic and semantic tagging of words in the transcripts, transcript segmentation, and lexical chaining.

### 4.1. Syntactic and Semantic Tagging

The first task in our approach is the syntactic and semantic tagging. This is carried out on the transcripts using WMATRIX which includes a set of linguistic tools namely CLAWS (part-of-speech tagger), SEMTAG (word-sense tagger) and LEMMINGS (a lemmatiser) as well as statistical functions such as frequency lists and KWIC concordances (Rayson 2001). In this paper we will concentrate on semantic (sem) and part of speech (pos) tags. An example of WMATRIX output is as shown below:

```
<w id="82.13" pos="NP1" sem="Z2">Stafford</w>  
<w id="82.14" pos="RR" sem="N6">once</w>
```

where 'w id' denotes the position of the word in the transcript e.g. the 13<sup>th</sup> word in the 82<sup>nd</sup> sentence. The pos tag 'NP1' denotes a singular proper noun, and 'RR' a general adverb. The sem tag 'Z2' denotes a geographical name, and 'N6' a frequency.

### 4.2. Transcript Segmentation

The process of extracting issues and decisions starts by breaking the original transcript into segments that address the same topic. A major issue in text segmentation is the choice of segment homogeneity criteria. Most segmentation algorithms use linguistic criteria such as cue phrases, punctuation marks, and references as reliable indicators of topic shifts (Kehagias *et al.* 2003). According to Halliday & Hasan (1976) and Barzilay and Elhadad (1997), parts of a text which have a similar vocabulary are likely to belong to a coherent topic segment. Due to the complexity posed by transcripts<sup>2</sup>, our approach to text segmentation begins with division of the transcript into blocks of fixed length, and then applying the lexical chaining technique to measure the cohesive strength between adjacent blocks which are then merged into meaningful segments. The lexical chaining approach is a linguistic technique, which clusters words into sets of semantically related concepts e.g. {coursework, assignment, assessment, examination} or into sets

---

<sup>2</sup> Transcripts, as a spoken language, pose a potential complexity due to their structure, sentence incompleteness, use of slang, ambiguity created when people use words that either address some politeness, uncertainty, or power relationship and references are made visual context.

with specific association such as hyponymy e.g. {IBM is a specialisation of a PC} or meronymy e.g. {hard disk is part of a PC}. The objective of the chain formation is to capture the topic and subtopic(s) discussed in the meetings, and track the corresponding actions. In the next section we describe in more detail the lexical chaining technique we have adopted in our segmentation task.

### 4.3. Lexical chaining

Cohesion, as introduced in Halliday and Hasan (1976), relates to the fact that the elements of a text ‘tend to hang together’ (Morris & Hirst 1991). Cohesion is achieved through the use of grammatical cohesion (i.e. references, substitution, ellipsis and conjunctions) and lexical cohesion (i.e. semantically related words). Lexical cohesion is used as a linguistic device for investigating the discourse structure of texts, and lexical chains have been found to be an adequate means of exposing this structure. Ellman & Tait (2000) see the lexical chaining approach to text analysis as highly attractive, since it is both robust and deals with the whole text. Lexical cohesion occurs not only between two terms, but among sequences of related words, called lexical chains (Morris & Hirst 1991). Boguraev & Neff (2000) point out that cohesion can best be explained by focusing on how lexical repetition is manifested, in numerous ways, across pairs of sentences. Repetition itself carries informational value, to the extent that it proves a reference point for interpreting what has changed (and thus what is at the focus of attention of the discourse). An example of this is shown in Figure 3.

*AAA:	If we can get some business done then. Agenda items <unclear>. How does the <b>project management</b> work, Mr. <b>Project</b> Manager?
*RRR:	Sorry. Well, as far as I know I’m due to travel around all over the country to various places <unclear>. It’s actually, has everybody seen the original <b>project</b> bid.
*BBB:	So according to the bid, 35 per cent of my time is <b>project management</b> . I think the idea is that I take the load off the principle investigators at the two sites, um, day to day <b>management</b> of the <b>project</b> . Sort of administration. type things, web site, the BSCW server as well. Was there a specific query on how the <b>project management</b> works?
*AAA:	No we know the principle of <b>project management</b> , but how it's going to work on this <b>project</b> .
*BBB:	So the <b>project management</b> issues we contact yourself, or through the principal investigators?

Figure 3. A portion taken from block 4 of transcript ‘A’

Figure 3 also displays the frequency of occurrence of a set of concepts in a set of blocks, and highlights a repetition of some of these concepts such as ‘*project management*’. This type of lexical repetition can, to a certain extent, be used to determine the degree of cohesion between adjacent blocks, and is capable of outlining issues raised at a certain time in the meeting. Repetition goes beyond the notion that discourse fragments with shared contents will also share vocabulary, the occurrence of repetition itself confers informational value which provides a reference for interpreting the focus of attention of the discourse and when the changes in that focus occur.

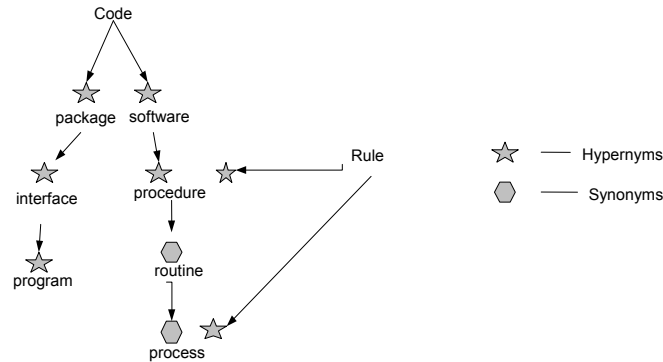
Table 1 illustrates the frequency distribution of the lexical chain which has the conceptual constituent<sup>3</sup> ‘*project management*’. In block 3 the issue on project management is introduced, it is developed further and becomes the main issue in block 4, and the issue is closed in block 5. As Phillips (1985) points out, the lexical inventory of a text is tightly organized in terms of collocation. It is this particular property that handles the overall organization of text, in general and on the identification of the *introduction of an issue*, *issue as a main discussion*, and *issue closure* at a particular time.

<sup>3</sup> Conceptual constituents are those in which the semantics of the whole cannot be deduced from the meanings of the individual constituents e.g. project management, management of project. In some texts they are classified as a type of collocation.

Concepts/Blocks	1	2	3	4	5	6	7	8	9	10	...	20	21
meeting	1	0	0	0	5	0	2	0	0	0	...	0	1
part	0	1	0	0	1	1	0	1	0	0	...	1	0
kit	0	0	2	0	0	0	0	0	2	0	...	0	0
project management	0	0	1	5	1	0	0	0	0	0	...	0	0

**Table 1.** The frequency distribution for ‘project management’ words.

In this experiment we have limited the primitive features used in the concept (chain) formation process to words that appear as lexical repetitions, synonyms or hyponyms. To start with, concepts were manually identified using Wordnet (Fellbaum *et al.* 2001), then automatically identified using WMATRIX. In Wordnet, nouns, verbs, adjectives, and adverbs are arranged into synsets (e.g. synonyms, hyponyms, meronyms) which are further organised into a set of lexical source files by syntactic category. Figure 4 shows an example of a lexical chain, a semantically related nominal group, in which each sense associated with a word has been expanded based on Wordnet. The figure shows a chain formation process indicating a relationship between the word ‘code’ and the word ‘rule’.

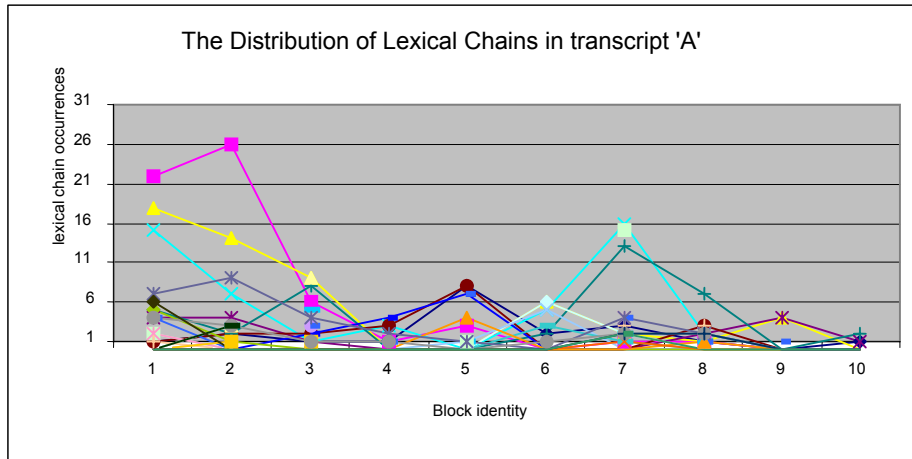


**Figure 4.** An example of a lexical chain.

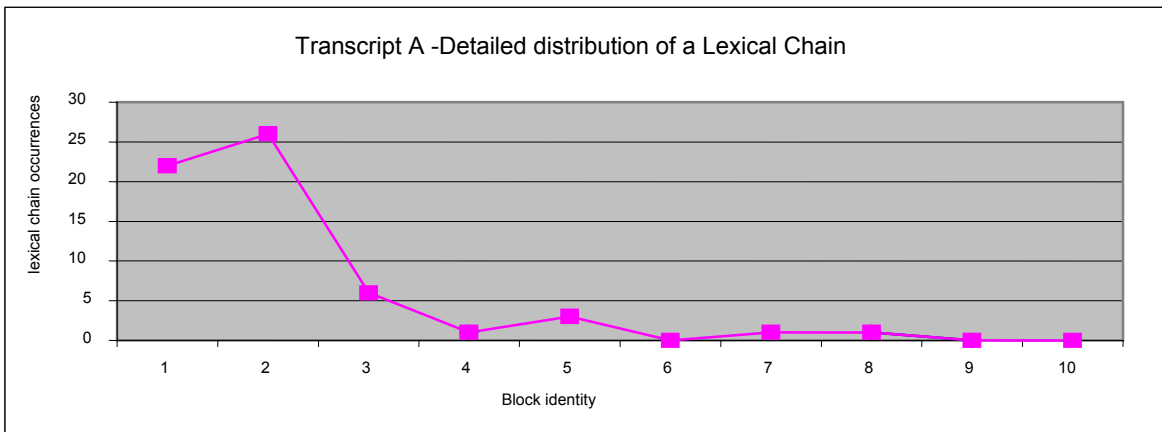
Our chaining algorithm starts by selecting all nouns in the transcripts, then computes their global frequency. Those nouns which appear below a specific threshold are considered to have less contribution to the topic and are ignored. Nouns with frequencies above the threshold are grouped semantically, yielding a series of lexical chains. Each word that is used in the formation of a chain must originate from a given block. All chains are weighted in terms of the frequency in which they occur in every block. If a chain is the most frequent in a particular block, then that chain carries the main issue which was discussed at that time. If the chain contains more than one lexical item the main issue is chosen by looking at the highest occurrence of the lexical item, while the other items of a chain are regarded as sub-issues or options. According to Green (1999), a point where several chains end and different chains begin is a good candidate for an issue boundary.

## 5. Experimental results

The lexical chaining algorithm was applied to transcript ‘A’ using Wordnet and WMATRIX. Informal analyses of the results were encouraging, suggesting that the lexical chaining algorithm can capture the majority of the issues that were covered in the meeting. Results for the analysis using Wordnet are shown in Figures 5 and 6 below.



**Figure 5.** Distribution of lexical chains within transcript ‘A’, using Wordnet



**Figure 6.** The life-cycle of a single lexical chain from transcript ‘A’

Figure 5 shows the distribution of lexical chains representing frequently discussed issues in transcript ‘A’. In Figure 6 one of the issues represented by the lexical chains shown in Figure 5 is presented in detail. This chain, which contains the lexical items ‘word’, ‘language’, ‘information’, ‘format’, ‘formatting’, and ‘morpheme’, can be seen as being the most frequently occurring lexical chain at the start of the meeting. The occurrence of the items in this chain decline in later parts of the meeting indicating the end of the discussion of the issue, this occurs between blocks 3 and 4.

Figure 7 shows the results for the analysis using WMATRIX. Graph ‘A’ presents the distribution of lexical items selected on the basis of high frequency occurrence in the whole transcript. Graphs ‘B’ and ‘C’ present lexical chains containing specific lexical items and demonstrate how changes in the distribution of the lexical items between blocks can be used to identify the occurrence of issues in this transcript.

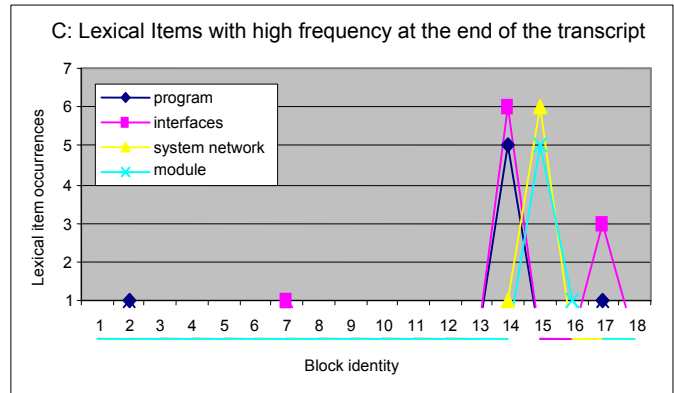
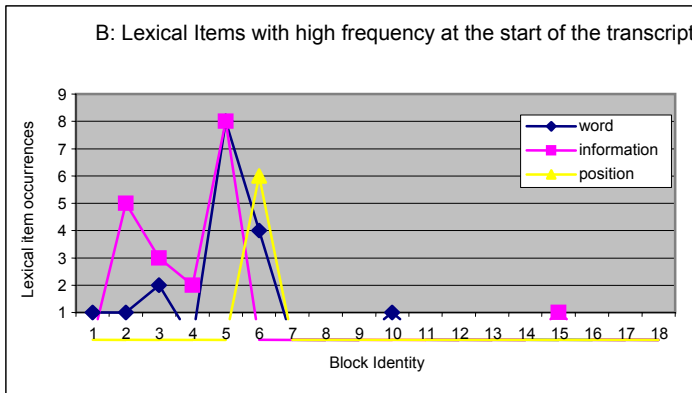
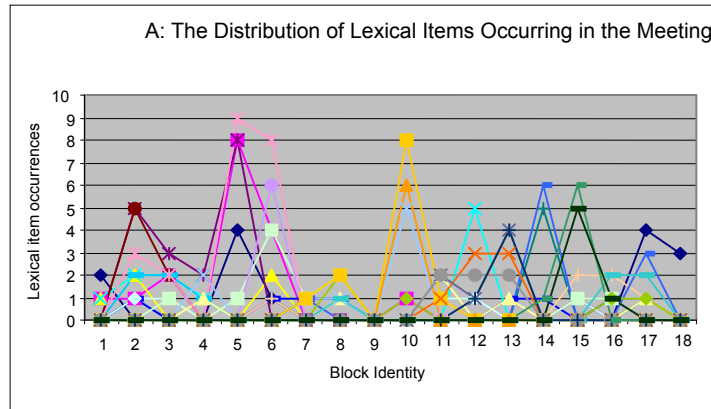


Figure 7: The distribution of lexical items within transcript ‘A’ using Wmatrix

Section 4.3 illustrated that every issue has three life-cycle phases: the introduction of an issue, the issue’s main discussion point (peak point), and the issue’s closure. Figure 7B shows that the introduction of the issue occurs in block 1, the main discussion occurs between block 2 and block 6 and the issue is closed in block 7. The area between one issue’s closure and the introduction of another issue (referred to as a *trough*) is crucial in our analysis because it allows us to extract decisions made after the closure of the previous discussed issue. Kazman (1995) argues that the changes in the temporal structure of a meeting always indicate decision points, even though these are often not recognised as such by the participants.

## 6. Current problems and future work

There are many factors that can affect the final word chain representation of a document, ranging from the nature of the chaining algorithm to the type of text used<sup>4</sup>. However the single biggest influence on the quality of the resultant lexical chaining is the knowledge source used in their creation. In other words the quality of our lexical chain formation is directly dependent on the comprehensiveness (or complexity) of the thesaurus used to create them. We used Wordnet and WMATRIX in our chaining process, and the following factors were identified as contributing to a reduction in the effectiveness of the disambiguation process and the comprehensiveness and accuracy of the final chain representation.

<sup>4</sup> structured, semi-structured or unstructured text.

### Wordnet factors:

1. Missing semantic links between related words.
2. Inconsistent semantic distances between different concepts.
3. Overloaded synsets such as 'being' which are connected to a large number of synsets. These types of synset cause spurious chaining, where an unrelated word is added to a chain based on a weak, yet semantically close relationship with one of these overloaded synsets.
4. No means of correlating the relationship between proper nouns and other noun phrases e.g {George Bush, US president}, which are crucial in representing event identity because they reflect the *who*, *what*, *where*, *when*, and *how* a decision was made in the meeting.
5. The level of sense granularity used to define word meanings is often too fine<sup>5</sup> for the chain formation process (Stokes et al. 2002).

The last two factors are particularly important when considering the similarity between documents and clusters in tracking issues, actions, and decisions made over a sequence of meetings.

### WMATRIX factors:

1. Some difficulties which are associated with the semantic tagging features and the limited domain.
2. WMATRIX operates on single words This makes it complex to extract issues represented by combined words like 'project management' (see figure 3) as the two words will be separated and probably allocated different frequencies.
3. Some semantic classification are very abstract and general, and this makes it difficult to extract the issues from a given text, e.g words 'sort of', 'kind of', can be used several times in the text as filler phrases, but WMATRIX will weight them as important nouns which are then incorrectly attributed as an issue.

## **7. Conclusion**

Through the ongoing experiments, and the findings obtained from the above analysis, we are now in a position to extract issues from the transcripts of meetings. We will also be able to use similar procedures to extract actions that are associated with each issue, and use the troughs to identify where decisions are made.

Upon completion of this research, we will be able to analyse cohesion indicators and use these indicators to summarise transcripts of meetings, then identify points in a narrative where decision elements emerge and alternate.

The main contribution of this paper is the introduction of a new text mining application within the decision management domain, whereby lexical chains are used to extract issues, as an indicator to actions and the associated decisions. We believe that giving users some means to interact with the records of stored meetings may alleviate potential problems associated with meetings. One potential problem that may be addressed is a reduction of rework. Other applications for this research include the knowledge management domain, particularly in large companies whereby the employees are often unaware that solutions for their problems may already have been determined in previous projects or other working teams. Currently, our method uses nouns to extract issues, in future research pronouns and other reference word groups will be incorporated in order to improve the process of issue extraction.

---

<sup>5</sup> e.g. the word 'city' in Wordnet is represented as:

An incorporated administrative district establish by a state charter  
A large densely populated municipality  
An urban centre.

When disambiguating a word like city for the chain formation this level of sense distinction is unnecessary.

## Acknowledgement

This work was conducted under the auspice of the TRACKER project, UK EPSRC grant (GR/R12183/01).

## 8. References

- Barzilay, R. & Elhadad, M. 1997, 'Using Lexical Chains for Text Summarization', in *Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97)*, Madrid, Spain.
- Boguraev, B.K. & Neff, M.S. 2000. Discourse segmentation in aid of document summarization, in *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, (HICSS) IEEE, Maui, Hawaii, pp.778 – 787.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. 2000, '*CRISP-DM 1.0 Step-by-step data mining guide*', SPSS Inc, USA.
- Ellman, J. & Tait, J. 2000, 'On the Generality of Thesaurally derived Lexical Link', in *Proceedings of JADT 2000, the 5th International Conference on the Statistical Analysis of Textual Data*, eds. M. Rajman & J.-C. Chappelier, Swiss Federal Institute of Technology, Lausanne, Switzerland. Ecole Polytechnique de Federale Lausanne, Switzerland, pp. 147-154.
- Feldman, R. & Hirsh, H. 1996, 'Mining Associations in Text in the Presence of Background Knowledge', in *Proceeding of the 2nd International Conference on Knowledge Discovery (KDD-96)*, Portland.
- Fellbaum, C., Palmer, M., Dang, H. T., Delfs, L. & Wolf, S. 2001, 'Manual and automatic semantic annotation with WordNet.', in *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh.
- Green, S. J. 1999, 'Lexical semantics and automatic hypertext construction', in *ACM Computing Surveys (CSUR)*, vol. 31, ACM Press New York, USA.
- Halliday, M. & Hasan, R. 1976, *Cohesion in English*, Longman, London.
- Hidalgo, J. M. G., Artificial, D. D. I., Politécnica, E. S. & Madrid, U. E. D. 2002, *Text Mining and Internet Content Filtering*, Available: [<http://www.esi.uem.es/~jmgomez/tutorials/ecm>], accessed Nov. 2003.
- Kazman, R., Hunt, W. & Mantei, M. 1995, 'Dynamic Meeting Annotation and Indexing', in *Proceedings of the 1995 Pacific Workshop on Distributed Multimedia Systems*, Honolulu, HI, pp.11-18.
- Kehagias, A., Fragkou, P. & Petridis, V. 2003, 'Linear Text Segmentation using a Dynamic Programming Algorithm', in *Proceedings of 10th Conference of European Association for Computational Linguistics*, Budapest.
- Liddy, E. 2000, 'Text Mining', *Bulletin of the American Society for Information Science*, vol. 27, no. 1, <http://www.asis.org/Bulletin/Oct-00/liddy.html>.
- Minneman, S. & Harrison, S. 1993, "'Where Were We: Making and Using Near-Synchronous, Pre-Narrative Video'", in *Proceeding of the ACM Conferense on Multimedia*, Toronto, pp. 207 - 214.
- Morris, J. & Hirst, G. 1991, 'Lexical Cohesion Computed', *Computational Linguistics*, vol. 17, no 1, pp. 21-48.
- Phillips, M. 1985, 'Aspects of Text Structure: an Investigation of the Lexical Organisation of Text', Elsevier Science Publishers, Amsterdam, Holland.
- Rayson, P. 2001, 'Wmatrix: a web-based corpus processing environment.' in *ICAME 2001 conference*, Université Catholique de Louvain, Belgium.
- Rayson, P., Sharp, B., Alderson, A., Cartmell, J., Chibelushi, C., Clarke, R., Dix, A., Onditi, V., Quek, A., Ramduny, D., Salter, A., Shah, H., Sommerville, I. & Windridge, P. 2003, 'Tracker: a framework to support reducing rework through decision management', in *Proceedings of 5th International Conference on Enterprise Information Systems ICEIS2003*, Angers - France, pp. 344 - 351.
- Salter, A., Shah, H. & Sharp, B. 2004, 'Reducing Rework in the Development of Information Systems Through the Components of Decisions', in *Proceedings of the 6th International Conference on Enterprise Information Systems (ICEIS2004)*, Portugal.
- Sardinha, T. B. 1995, 'The OJ Simpson trial: Connectivity and Consistence', in *BAAL Annual meeting*, Southampton, UK.
- Stokes, N., Carthy, J. & Smeaton, A. F. 2002, 'Segmenting Broadcast News Streams using Lexical Chaining', in *STAIRS 2002*, Lyons, France, pp. 145-154.
- Tan, A. H. 1999, 'Text mining: The state of the art and the challenges', in *Proceedings of the Pacific Asia Conf on Knowledge Discovery and Data Mining PAKDD'99 workshop on Knowledge Discovery from Advanced Databases*, pp. 65-70.
- Weeber, M., Vos, R. & Klein, H. 2001, 'Using concepts in literature-based discovery: Simulating Swanson's Raynaud-Fish oil and Migraine-Magnesium discoveries', *Journal of American Society for Information Science and Technology*, vol. 52, no. 8, pp. 548-557.
- Whittaker, P., Hyland, P. & Wiley, M. 1994, 'Filochart: Handwritten notes Provide access to Recorded Conversations', in *Proceeding of the ACM Conference on Human Factors in Computing Systems (CHI'94)*, Boston, pp. 271 - 277.