Data Refining for Text Mining Process in Aviation Safety Data

Olli Sjöblom

Turku School of Economics, Rehtorinpellonkatu 3, 20500 Turku, Finland

Abstract. Successful data mining is an iterative process during which data will be refined and adjusted to achieve more accurate mining results. Most important tools in the text mining context are list of stop words and list of synonyms. The size and richness of the lists mentioned depend on the structure of the language used in the text to be mined. English, for example, is an "easy" language for search technologies, because with a couple of exceptions, the stem of the word is not conjugated and terms are formed using several words instead of creating compounds. This requires special attention to definitions when processing morphologically rich languages like Finnish. This chapter introduces the need and realisation of refining the source data for a successful data mining process based onto the results achieved from first mining round.

Keywords: Data Mining, Text Mining, Flight Safety.

1 Introduction

Air transport is among the safest modes of transport. However, although the global rate of accidents is stabilising and the situation is somehow satisfactory, the growth in air traffic will increase the absolute number of fatal accidents per year and therefore new ways of improving air safety need to be explored [3].

Therefore the newest research resources and data processing techniques are unavoidable. The focal point in this context is analysis. Fundamental to every Safety Management System is the principle of collecting and analysing operational data [5]. The only way to process narrative data with computing has till the end of the 1990s been to utilise query tools for specific issues of data base systems relying on the skills and experiences as well as the memory of the safety officer [13]. The idea to apply data mining and especially text mining as investigation and analysis tool for flight safety reports was created by the author of this article when he worked as inspector in Investigation and analysis unit in Finnish Civil Aviation Authority (FCAA).

2 Data Mining

A data mining system cannot be described easily, because it can use several of these methods as a combination. A rather good definition among many of them is that by

Parsaye who regards data mining as *"a decision support process in which we search for patterns of information in data"* [14]. Simply expressed, the goal of data mining process is extracting high-level knowledge from low-level data. That is to search for relationships and global patterns that exist in large databases but are *hidden* in the vast amounts of data, which might take years to find with conventional techniques [18]. The significant development in database and software technologies, i.e., warehousing of transaction data has enabled the organisations to build the foundation for knowledge discovery in databases (KDD) which consists of such phases as selection, pre-processing, transformation, data mining and interpretation/evaluation [1].

The data mining process contains several steps or phases which chain from data to knowledge is presented in Figure 1. The first results of data mining process after representing the discovered patterns need to be evaluated carefully by consolidating them with the existing domain knowledge, which is more a combination of art and common sense than science [8]. Discovery process often begins with making a hypothesis, but using data mining this is not necessary. Text mining, that this article is about, is one subclass of data mining [19].

Usually the data accumulates in organisations faster than it can be processed [17]. The need for automated means to process the data is also increasing rapidly [2]. Skilled analytical and technical specialists are still required, because data mining process does not give straight answer to questions. Its role is purely a decision support system [10]. When used by professionals, data mining can discover patterns and anomalies that have not been thought before [13]. The main target of data mining in aviation safety is to find hazardous trends and patterns in collected flight safety data. Despite of a couple of successful projects, the volume of using data mining is at a very moderate level. As a bottom line these projects produced a statement: "Very encouraging and very promising; more work to be done" [12].

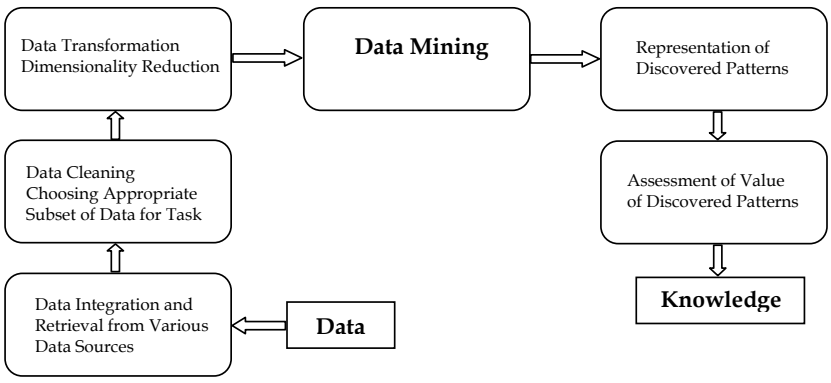

Fig. 1. The Knowledge Discovery in Database process [4], [6]

The structured data can be analysed easily using queries from databases and running their results through graphic tools but there have not been tools to analyse textual data before text mining. It is estimated that the main part (over 80%) of the information in reporting is written in the unstructured and textual format. Narrative text mining is demanding because of the multiplicity of languages spoken in the world.

2.1 Data Mining in Finnish

The Finnish language is a compound rich and very inflectional language having an inflectional and derivational morphology. As English has only two, there are 15 cases in Finnish. Finnish case endings correspond to prepositions or postpositions in other languages (cf. ilma/an, ilma/ssa, ilma/sta, in English to the air, in the air, from the air) [7]. In Finnish, a noun may have some 2,000 forms, an adjective 6,000, and a verb 12,000 forms. If the effect of derivation is also taken into account, the figures will be increased roughly by a factor of 10 [9].

The FCAA has used “normal” tools like Excel and report generators of database systems lacking a tool that could process safety report narratives. Using flight safety reports as test material text mining was tested. The study can be considered as a jump to the unknown, because, as far as the authors’ knowledge, text mining in Finnish has been generally made only to some extent and nobody has applied text mining to Finnish flight safety reports before.

2.2 Research Process

The research process was planned to follow the process presented in Figure 1. As the first phase, flight safety reports from years 1994-1996, about 1200 cases, were chosen. All the test data was extracted from one Microsoft Access database in .csv form. Taking the narratives only for final test material can be regarded as phase two, choosing appropriate subset of data. The cleaning process, in this case correcting misspelled words and expanding acronyms, happened through creating list of synonyms. As for the third phase, choosing narratives can be seen as dimensionality reduction and due to source, no data transformation was needed. Using complex mining techniques, data needs to be cleaned frequently from inaccuracies like duplicate records, anomalous data points and human errors [15].

Three different systems seemed to be appropriate enough for comparing the expected results as benchmarking. The author was aware of one prototype and one commercial product with Finnish module prototype as well as one commercial system with encouraging results mining Spanish, which seemed worth testing Finnish.

The processing systems are normally built to recognise formal language only. This requires a list of words that the system does not recognise. Stop words list includes words that will be ignored. These are the ones that do not provide any insight, like: a, about, above, across, after, almost, although, always, etc.

Stop words list was created manually using the filtered list of words. From the total amount of more than 8000 filtered words 446 stop words were listed. The verb ‘to be’ in all its forms was defined as a stop word, because it appears rather often in the reports and it can freely be ignored without losing any information.

List of synonyms is mapping multiple synonyms to one word including misspellings and abbreviations. Optionally, non-alphabet and non-number characters are mapped to ‘blank’, and numbers are mapped to ‘delete’. Some examples of these are:

- " <blank>
- ` <blank>
- ' <delete>
- a/c aircraft
- apparent appear
- apparently appear
- kts knots
- rwy runway
- taxied taxi
- taxiing taxi

The list of synonyms, like stop words, was created manually from filtered list of words and it contained 2048 words, a rather high amount because of the characteristics of Finnish.

3 Evaluation of the Tools

Clustering was chosen to be used, because the main target especially was to find similar cases from a vast amount of report data without having any presumption whether these exist or not. The length of the reports in the basic data varied remarkably. Therefore, the question about data quality that means data accuracy and completeness emerges as one of the biggest challenges for data mining. before.

3.1 Tool #1

The Tool #1 is primarily from Xerox laboratories, America, but now developed and represented by Text Mining Solutions, TEMIS, in France. The Value-Added Reseller for TEMIS products in Scandinavian is Lingsoft Oy in Turku, Finland [11]. Testing this system was prototyping, because the module for processing Finnish was quite new and there were no experiences in a wider range.

TEMIS decided to create 26 clusters leaving out 8 reports as unclassified documents, as which system handled each report. Their size varied from 109 to 21, and due to mining rules the biggest cluster was divided into two sub clusters with 58 and 50 documents. Every document is listed with its percentage explaining the cluster, the highest being 27 which be regarded significant in this context. The

smallest clusters produced some applicable information. However, although the new Finnish module worked properly, a list of special synonyms and acronyms depending on the context might be needed because they do not exist in the thesaurus of the mining system.

3.2 Tool #2

The tool #2 is developed at Tampere Technical University, Signal Processing Laboratory as one of the results of the project called GILTA (manaGIng Large Text mAsses) between 1999 and 2003. The scientific objective is to find out if it is possible to find the desired things in a mass of documents by using self-organising maps (SOM) or similar methods combined with linguistic sentence analysis. Simply expressed, the documents are converted into vector form. The vectors determine the similarity of the documents through using Euclidean distance. The shorter the distance, the more similar the documents are. This method provides total language independence. Using systems training the methodology can easily be adapted to any application field [16].

The system itself decided to share the data into 100 clusters when the amount of most significant words was chosen to be 9. The size of the clusters varied between 158 and 1. The amount of the clusters can be considered to be appropriate in this context, but the distribution of reports should be more uniform.

3.3 Tool #3

The tool #3 is a commercial product of American origin called PolyVista. The company is located in Houston, Texas. It is a full-capable data mining system with several components and technologies built-in having ability for mining both numerical and text data. Also with this tool only clustering was performed using different amounts of clusters that could be determined by the user. This was the first time when Finnish text was mined. The tool is built to be used in English context, but tests had been made with Spanish achieved good and encouraging results and that is why the company dared to test its applicability for Finnish. This seemed to be successful as well.

The results were returned to the researcher in form of pictures taken from the Cluster Browser window. The first run was kind of a preliminary one made without any pre-processing and determining 15 as the amount of clusters by a guess. After using stop words the results looked different and the process was continued by taking synonyms with. The data was processed determining the number of clusters first to be 6 and then raising it up to 20 with step of 2. All the research phases in the process of estimating this tool have produced some kind of good and applicable results, especially when the smallest clusters are observed. The scores of most important words are high and quickly estimated the words seem to be of significant character.

4 Preparation of the Test Material for the Second Round

The first results proved an obvious need for tuning the test material, especially the definition of stop words and synonyms. This was seen through the mining results of GILTA system, because they were listed in a clear form of Excel datasheet.

The search technologies are often confronted by great challenges in conjugation forms and compounds. In Finnish, for example, the words may have thousands of conjugation forms and in addition to that, they can be parts of compounds in almost countless amounts. In average, every seventh word can be found in its basic form in fluent Finnish texts. For search technologies English is an "easy" language. With a couple of exceptions, the stem of the word is not conjugated and instead of compounds, terms are formed using several words. Because in most of the information systems used search technologies are constructed onto English, the conjugated forms of compounds are often not found. This happens especially in processing large amounts of documentation in morphologically rich languages like Finnish.

If the mining system contains language tools, like TEMIS, there is generally no significant need for defining words. When processing short texts including special terminology there is. As an example can be taken the words 'gear' and 'landing gear (one compound in Finnish)', as well in English as Finnish, which are pure synonyms in this context, but not necessarily in general language.

It was a pure mistake, that one of the most common stop words '*jälkeen*' (after) was forgotten from its list. The consequences were clear in its frequent appearance in the most significant words of the clusters. These kinds of mistakes are worthy to correct, because the word mentioned appears 304 times totally. Despite of its characteristics of an obvious stop word, its frequency in this size of data is significant, which would anyway require paying special attention onto it. Another obvious mistake was that some tenses were not taken into account when list of synonyms was created.

One of the most remarkable problems in this study appeared due to the high frequency of the word '*kone*' (plane). This caused the class #11 to contain 158 reports, which makes no sense, at least in this test data context, where the appropriate cluster size should be at most about 25 reports. The word appeared among the nine most significant words in 42 classes and as the most significant word in 5 classes.

What to do with a word that causes the enormous size of one cluster but which is among the most important if not even one of the really most important words? The first idea would not be just to ignore it, and that is why it required careful analysing its role in the data. This problem acting as an efficient accelerator, quantitative data analysis application called NVivo was taken into use to get deeper analysis information. This move appeared to be a very successful one, because using Excel and Word alone would not have led to a satisfactory goal in understanding the relations between the concepts in test data.

The first test with using NVivo was a simple query which had the definitions for searching all words in different forms meaning aeroplane, so, text search criteria were (in English) plane OR aeroplane OR aircraft OR a/c. Although Finnish is a rather complicated language, using wildcard – any characters (*) after the stem of the word, all forms were taken into the result of the query, because all forms of this word in this

context, according to the list of words, do not make changes to its stem. NVivo takes into account all the words defined in the query options and determines their relevance in the data mass. GILTA counts the word only once if it can be found in the report although would appear several times. This causes differences when distributions of words and their significances are estimated in this research process. NVivo counted 67 classes including word ‘kone’ its relevance varying between 34% and 1% as GILTA has taken 42 classes into account. The word ‘kone’ (plane) has a total of 680 appearances in all its case forms. At the first round, its synonym ‘lentokone’ (airplane) with 51 appearances was taken into account, but ‘ilma-alus’ (aircraft), 20 appearances, was kept separate, although it means the same thing. Aircraft can, of course, have several meanings (e.g., helicopter, glider, etc.), but in this context there is practically no difference between these concepts, especially as its usage in the reports is almost nil which can be seen from its frequency in the data, too. However, the English abbreviation a/c with total of only three appearances had been added to the synonyms of ‘kone’ (plane), which brought one report to class 11, in which it appeared two times. Combining ‘plane’ and ‘airplane’ raised the amount of the concept from 680 to 731, only about eight per cent. Adding the word ‘aircraft’ to the synonyms would increase the number by less than three per cent, so the latter can be seen not having any effect to the final results, although, as mentioned before, it caused one more report to belong to the class 11.

Class 34 must be mentioned separately, as being rather well-formed and homogenous, because it contains reports written about crossing border without permission. It will be interesting to see during the second round how this will change because the frequencies of the expressions ‘without permission (24)’ and ‘plane (22)’ are so close to each other. Speculatively scanned, ignoring the word ‘plane’ would not seem to change the core information of the class.

During the process of analysing how this word possibly affects the results, it was noticed that this word was found in the most important words 454 times, so, the rest 297, close to the half of the whole amount, stayed beyond the significant words.

The further this special problem was studied the cleared was noticed that from the point of view of the whole study it really makes sense to solve this basic seeming question very carefully. In the final test the query definitions included all the different forms and synonyms of the word were taken into account. It appeared in 45 classes with its relevance varying between 4.9 and 1.0 per cent. The four highest had their relevance between 4.0 and 4.9, the next four between 2.1 and 3.3, and the rest having less than 2.0, among which the relevance of the last 14 classes was exactly 1.0%.

This result gave a slight presumption to ignore the word in mining process. Because there were no more than those 45 classes mentioned, it was simple to manually go through all the classes. This investigation resulted to the finding that the word can very obviously be added to the list of stop words, because it could have been ignored in the reports without losing relevant information in them. With this manual check the results achieved from using NVivo were confirmed.

All the airfield abbreviations for airports and airfields in Finland (EF++ which stands for Europe - Finland + two characters for the place name, for example EFTU = Turku airport), were removed. These were 141 different expressions, one abbreviation alone or a combination of two or more abbreviations describing for example the route, making together 363 ‘words’. This was done because the place, in a small country

like Finland, does not contain information value in mining process, considering the uneven distribution of especially the airline traffic which concentrates into Helsinki. From the mining results it could be easy to draw the conclusion that Helsinki is the most dangerous airport in Finland with large marginal to others, which is not the case, indeed. If the place has any significance to the mining process, it will surely be discovered through other deviations during the deeper analysis of the mining results.

Helsinki as a name can be found 103 times in different forms in the data, being the most significant word with 24 appearances in class 51 that contains 24 reports. Interesting point in this class is that the next important word with 18 exemplars is *to return*. In this class the most cases are about returning to the airport because of some reason and it can be stated that the airport does not have any significance, it just happens to be Helsinki because of the traffic volume. It does not appear in any other class among the nine most significant words, so it could be added to the stop words. In case something happens especially at or close to a certain airfield or airport, the possible place causing problems will certainly be found as one of the findings when the cases concerning the problem are examined more minutely.

Due to the same reason, *Finnair* with its abbreviations (together 92 words in six different forms) was removed, because of the multiple volume of traffic compared with the total of the other operators together. It is worth noticing that it was the most significant word with frequency of 15 in class 40, which contained 17 reports. Generally, what happened and for which reason is more important than to whom it happened, which can be as well derived from the reports in case it would be important. This is on view very clear in this class also. A special attraction was focused onto class 40, where the two most significant words were ‘Finnair’ (or FIN) and ‘kone’. It is worth noticing that in this case the conjugation forms and cases only of the word are taken into account, not its synonyms. This combination produced the relevance of 42.5% to the class mentioned. The next classes were 11 with relevance of 25.8%, 82 with 11.7%, 4 with 10.5%, and 78 with 10.1% the rest 22 classes with their relevance varying from 9.9% till 2.2%. In different contexts, however, these kinds of removals could not have been done, because bigger countries even in Europe have several air traffic hubs and big companies, in which case the company and the place might play an important role in report analysis.

Another significant discovery that formed a bit complicated situation was the rather important concept *Flight Level* and especially its abbreviation, *FL*, because in this process could be stated that ‘everything depends on everything’. It figured 50 times which brought it to the class of rather frequent words and thus its impact should be decreased. When the written expressions for it with all synonyms and case forms are counted, the total amount will moderately exceed 150.

General overview showed that it has a significant frequency only in class 22 in which it was the most describing word with frequency of 12. The size of this class is 12 reports and the abbreviation can be seen separately 26 times and five times written together with the numbers (for example FL220, that means an altitude of 22 000 feet). The same abbreviation was available the second time only in class 10 as the last significant, the ninth, word with frequency of two (relevance 1.5%). This class contains seven reports of which the two most significant words have the frequency of six and for the third one the correspondent value is five. The analysis using NVivo confirmed the discovery, because in the class 22 the relevance of the abbreviation was

33%, but decreased beyond this class dramatically being between 2.5% and 1.2% in the remaining 13 classes in which it appeared.

It is, however, worth noticing that the criteria of NVivo differ from the correspondent of GILTA. This can clearly be seen in the query results of this special word, in which the most relevant matches with GILTA, but between the two classes GILTA has taken into account mentioned above, NVivo has put four classes with their relevance varying between 2.5 and 1.6. The same difference can be discovered from the query results of the synonyms. This discovery could cause somewhat confusion, but because the results are rather close to each other, it can be stated that accuracy achieved with these arrangements can be considered as sufficient taken into account the fact that all these results are preliminary and only leading the mining process further.

The impact of its synonyms, that means in this context, the words 'lentopinta (flight level)' or 'pinta (level), seems to be somewhat similar. The first one is the most describing word in class 42 with nine occurrences, the same number as class size. The same appears another time in class 22, where it is the least significant word with three occurrences. As for the relevance, there is class 48 with percentage of 11.7 between these two. The second one, 'pinta (level)' is distributed as presented in Table 1.

Table 1. Distribution of word 'pinta (level)'

Class #	Reports in class	Significance (9-1)	Number of appearances	Relevance %
48	11	9	7	6.2
92	4	7	2	4.0
24	2	1	1	2.3
42	9	5	3	2.1

NVivo has found together 17 classes including the word mentioned and given slightly different distribution compared with GILTA. The relevancies vary between 6.2 and 1.4 but the four classes in table can be found in top-eight, which proves that compared to the amount of classes, the results are not far away from each other.

If both synonyms with their different forms are combined, the results look as presented in Table 2. As seen from the two separate tables, one class might include both synonyms, which makes the situation slightly different. Number of appearances can be combined using the same pattern as GILTA (=in how many reports the word appears), but significance should be ignored, because counting them together or taking average does not make any sense in this context.

Table 2. Distribution of combined synonyms

Class #	Reports in class	Significance (9-1)	Number of appearances	Relevance %
42	9	9	9	32.3
48	11	9	7	10.8
22	9	5	3	5.0
92	4	7	2	2.6
24	2	1	1	1.6

Query for only synonyms mentioned before brought a list of 25 classes in which the relevance varied between 32.3% and 1.1% but the distribution was rather clear which can be seen also from the table. All classes in the table are among the 11 highest ranked by NVivo.

This case generally revealed the mistake that synonyms were not taken into account sufficiently because the abbreviation with the explicit words 'lentopinta (flight level)' and 'pinta (level)' were not combined to mean the same concept although case forms of both were returned their basic forms. All these observations prove that the concept Flight Level in its different forms cause distortion to mining results enough to be eliminated, that is to be put into stop words. This is grounded on accurate manual analysis of reports in clusters, where the word has significant meaning.

The next problem was caused by abbreviation of *foot* or *feet*, ft. It is a useful and common term and abbreviation in air traffic, but its role and significance in this context should be estimated again. It appears 87 times alone and once combined with / and /min. It is the most significant 'word' in class 82, where it appears in all 26 reports with total amount of 41 times. Carefully inspected it can be stated, that none of them has any significant contribution to the report primarily. For example, six reports have been written due to bird strike, in which case the most important point of view are the conditions and the phase if flight. This can primarily be analysed on basis of other information than altitude of the occurrence that could be of greater importance later. As for the appearance beyond class 82, the abbreviation ft is displayed only twice in the clustering results of GILTA, three times in one cluster as the last significant word and two times in another as the second last important. Thus, from the total number of 87, about half, 46 are counted into the most significant words, so, according to its distribution, the affect of this abbreviation can easily be removed from the data by adding it into stop words.

The English word *'gear'* (seven appearances) was added to its Finnish translation 'laskuteline' which was forgotten to do before. This mistake led to the discovery of another, as the synonyms 'teline (gear)' and 'laskuteline (landing gear)' were kept as separated words, which mistake surely affects significantly the mining results. The first word mentioned has 62 appearances and the second 57 which makes the total of 119, which amount can be considered as significant.

The word *route, 'reitti'*, appears in its different forms not less than 234 times, being the most significant word in four classes and among the nine most significant words in 26 other classes. This observation is in line with the test made with NVivo which found the word in 46 classes its relevance varying between 13.3 and 1.9 per cent. The high frequency might cause the ignorance of the word, but its distribution is so even that it cannot be expected to cause any skewed impact onto the clustering.

To *declare, 'ilmoittaa'* with its total 98 appearances in different case forms and tenses is an interesting piece of testing, because the verb with its forms was left out from synonyms list during the first round, although the synonyms of a correspondent noun, declaration, 'ilmoitus' was put onto the list. As a verb it appears only two times in the most significant words being the sixth with two appearances in class 63, the size of which is six reports, and in the similar way but with three appearances in the class 88, and according to the query performed using NVivo, the verb appears in 37 classes its relevance varied between 9.2 and 1.2 per cent. On the basis of this, it

cannot be considered to be significant at all. It is, however, worth to add to the synonyms list, because in this amount of test data, 98 appearances might change the distribution of classes. These both words might be ignored totally after having the results of the second round, because the most important thing is that something has happened and not that somebody declared something to be happened. As a general observation in the test data, the relevant information can be achieved from the reports leaving out the expression about declaration, but this has to be examined carefully.

Almost one hundred checking procedures were made with synonyms and stop words. After careful estimation about the impact of possible changes, no major ones were made, only the most obvious tunings in case forms which does not change the process but make the results to be more accurate. As a total, after this process, the list of stop words contained 569 words when 124 new words were added to it. The list of synonyms contained 2045 word definitions for the first round and it was changed slightly, when 74 new definitions were added, 25 were changed and 57 words were moved from this list to the stop words list, the total of definitions being 2062 for round two.

One essential point in this context was to keep the data as untouched as possible to keep the impacts of the changes as measurable as possible. The main purpose of this refinement is to correct the obvious mistakes, which include both evident mistakes, like forgotten values as well as consequential ones, like wrong estimations of words as for their role and impact to mining results. After the second results the data can be more refined and especially fine tuned, if and when the corrections can be proven to have been successful in the mining process.

5 Conclusions

Despite of the big amount of work to be done when analysing carefully the first results to refine the data for reaching better mining, the results of the first round were encouraging to develop the study further. According to them the direction is correct. The aim of the researcher was already at the very beginning not to stop after the first round but to continue the process and achieve more accurate and usable results. This was both because of the theory which, clearly states that iterations are normal, as well as the increasing interest to the research and its possibilities. All the tested system confirmed the fact that data mining and especially text mining should be an iterative process.

When this article was written, the refined lists were delivered to the operators for the second round. As these are processed, the research will continue by analysing the results carefully. This analysis will reveal whether the research has gone to the right direction or not, that is, the clusters contain more accurate information as for their size and content. However, all the time it should be taken into consideration that data mining tools are “only” decision support systems not giving straight answers to questions but they can well be utilised in scarce retrieval of information.

References

1. Blake, M. B., L. Singh, et al.:. "A Component-Based Data Management and Knowledge Discovery Framework for Aviation Studies." International Journal of Technology and Web Engineering **1**(1) (2006)

2. Delen, D., M. D. Crossland: "Seeding the survey and analysis of research literature with text mining." Expert Systems with Applications: An International Journal **34**(3), 1707-1720 (2008)

3. European Commission: Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on occurrence reporting in civil aviation. Commission of the European Communities. Brussels (2000)

4. Fayyad, U., G. Piatetsky-Shapiro, et al.: Knowledge Discovery and Data Mining: Towards a Unifying Framework. Second International Conference on Knowledge History and Data Mining (KDD-96), Portland, Oregon, AAAI Press (1996)

5. GAIN Working Group B: Role of Analytical Tools in Airline Flight Safety Management Systems, Global Aviation Information Network (2004)

6. Han, J., M. Kamber.: Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers (2001)

7. Karlsson, F.: Finnish grammar. Porvoo, WSOY (1987)

8. Kloptchenko, A.: Text Mining Based on the Prototype Matching Method. Turku Centre for Computer Science. Turku, Åbo Akademi University**:** 117 plus additional pages including original papers (2003)

9. Koskenniemi, K.: An application of the two-level model to Finnish. Computational morphosyntax: Report on research 1981-84. Helsinki, Department of General Linguistics (1985)

10. Kutais, B. G.: (Ed.) Focus on the Internet, Nova Science Publishers, Inc. (2006)

11. Lingsoft and TEMIS Announce Partnership to Expand Text Mining Coverage to Northern European languages and countries, http://www.lingsoft.fi/news/2005/temis.html

12. Muir, A.: Fundamentals of Data and Text Mining. Seventh GAIN World Conference, Montreal, Canada, Global Aviation Information Network (2004)

13. Nazeri, Z.: Application of Aviation Safety Data Mining Workbench at American Airlines. Proof-of-Concept Demonstration of Data and Text Mining. McLean, Virginia, US, Center for Advanced Aviation Systems Development, MITRE Corporation Inc. (2003)

14. Parsaye, K.: "A Characterization of Data Mining Technologies and Processes." Journal of Data Warehousing **2**(3), 2-15 (1997)

15. Seifert, J. W.: Data Mining: An Overview. Focus on the Internet. B. G. Kutais (ed.), Nova Science Publishers, Inc. (2006)

16. Toivonen, J., A. Visa, et al.: Prototype Based Information Retrieval in Multilanguage Bibles. WIAMIS 2001 (2001)

17. Wang, X., S. Huang, et al.: LSSVM with Fuzzy Pre-processing Model Based Aero Engine Data Mining Technology. Advanced Data Mining and Applications. Heidelberg, Springer Berlin / Heidelberg (2007)

18. Watson, R. T.: Data Management: Databases and Organizations, John Wiley & Sons (1999)

19. Visa, A., J. Toivonen, et al.: Data mining of text as a tool in authorship attribution. Data Mining and Knowledge Discovery: Theory, Tools and Technology III, Orlando, USA, SPIE-The International Society for Optical Engineering (2001)