# Using latent semantic analysis of email to detect change in social groups

Ian McCulloh, Eric Daimler, Kathleen M. Carley

***Abstract*—Email text data is a rich resource that, when properly used, may enhance warning of economically material events in commercial enterprises. Armed with the temporal text classification of an email time stamp on a large data set, we combine latent semantic analysis with statistical process control, to detect potential change in sentiment. A novel approach to identifying the causes of change is proposed by averaging concept scores across statistically significant time periods and then performing an inverse singular valued decomposition. The resulting term by significant time period matrix, is used to explore potential causes of change and are compared to historical events. Our findings suggest that causes for significant change in semantic intent can be identified in some circumstances. On a dataset of 50,000 unique emails from Enron, we demonstrate this novel approach to investigate semantic change in email text data.**

***Index Terms*— CUSUM, Enron, Latent Semantic Analysis, Statistical Process Control, Text Analysis, Text Classification.**

## I. INTRODUCTION

Change detection in semantic content represents an exciting new area of research. The area of statistical process control and quality engineering is as well established in academia as data mining and text analysis. The combination of these disciplines is likely to produce significant insight into organizational and communication behavior. Immediate applications to management are obvious.

Semantic change detection is essentially a statistical approach for detecting small persistent changes in the semantic content of an organization's communication over time. Organizations are not static, and over time their structure, composition, and patterns of communication may change. These changes may occur quickly, such as when a corporation restructures, but they often happen gradually, as the organization responds to environmental pressures, or individual roles expand or contract. Often, these gradual changes reflect a fundamental qualitative shift in an organization, and may precede other indicators of change. It is important to note, however, that a certain degree of change is expected in the normal course of unchanging corpora, reflecting normal day-to-day, text-to-text variability. The challenge of semantic change detection is whether metrics can be developed to detect signals of meaningful change in text in a background of normal variability.

Text data is available in many forms. One example is the available text on the web. With no search engine indexing more than about 16% of the estimated size of the publicly indexable web [1], at least 20 billion web objects have been indexed [2]. Approaches to classification have been developed to specifically be more effective in all-text environments [3].

Since becoming available in 2002, the Enron email corpus has been the subject of repeated study by researchers [4,5,6,7]. This dataset will likely continue to be of interest for some time; the corpus represents a body of correspondence within a defined social group at a scale unusual in the genre. We are fortunate to not only have this substantial collection of emails but also time stamps on them.

Timestamps on correspondence provide temporal richness to text classification. Tracking changes in the text classifications over time may highlight changes in the social structure as defined by the messages. This is not the study of sentiment per se, but rather of changes in sentiment. Rather than looking at what the sentiment tells us directly, we are looking at what the changes in sentiment tells us. We are concerned with the degree to which a change in sentiment can help to predict real world events. The ability to detect changes in social groups is important in a variety of applications.

This paper proposes a novel method combing statistical process control with latent semantic analysis to identify changes in the semantic content of an organization's text communication. A methodology for identifying the cause of change is laid out. The methodology is demonstrated on the Enron e-mail corpus.

## II. BACKGROUND

Applying classification algorithms and Latent Semantic Analysis (LSA) has recently been explored as a means of

This work was supported by the Center for Computational Analysis of Social and Organizational Systems, School of Computer Science, Carnegie Mellon University, http://www.casos.cs.cmu.edu.

I. McCulloh is with the Network Science Center, U.S. Military Academy, West Point, NY 10996 (phone: 845-702-9115, fax: 845-938-2409, email: imccullo@cs.cmu.edu)

E. Daimler is with Carnegie Mellon University, Building 23, Moffett Field, California 94035 (phone: 408-241-0055 email: edaimler@cs.cmu.edu)

K.M. Carley is with the Center for Computational Analysis of Social and Organizational Systems, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213 USA (phone: 412-268-8163 email: kathleen.carley@cs.cmu.edu)

detecting salient issues in text [8]. Approaches to change detection, such as statistical process control may be applied to quantifiable metrics from a time-stamped text data set. Change in this context may provide insight into the degree to which changes in salient words may be detected.

The classification of documents has been studied for texts in analog form, electronic form, emails, and even the Enron Corpus in particular [5,6,7]. These studies have ranged from inquiry into the effectiveness of document classification [4] to exploratory data analysis [8]. New research in text classification algorithms have been developed, and tested against existing corpora [9]. Newly collected corpora have been explored using existing classification techniques [4]. Of particular interest to this body of work, efforts have been made to classify online and email communication [8,9,10]. Classification algorithms have even been applied to Enron email communications as a test of the algorithms effectiveness [4].

There is evidence to suggest that TF-IDF may be an accurate method of estimating semantic content of the text sources mentioned above. Experiments on Web pages and TV closed captions demonstrate high classification accuracy for TFIDF [11]. While TFIDF has been explored [11,12,13], combinations of classifiers such as TFIDF with LSA [14] and K-Nearest Neighbor with LSA [15] has also been explored. This study suggests a method to further reduce the immense dimensionality of over 50K emails through the application of TFIDF with LSA, but then also applying a statistical process control chart from quality engineering to identify possible changes for investigation.

Statistical process control is a technique used by quality engineers to monitor industrial processes. They use control charts to detect changes in the mean of the industrial process by taking periodic samples of the product and tracking the results against a control limit. Once a change has been detected, the engineers determine the most likely time the change occurred so that they can reexamine and reset the process to avoid financial loss for the company by making substandard or wasteful product. Control charts are usually optimized for their processes to increase their sensitivity for detecting changes, while minimizing the number of false alarms – signals when no change has actually occurred in the process.

The control chart investigated for this study was the cumulative sum (CUSUM). The CUSUM control chart is a widely used control chart derived from the sequential probability ratio test (SPRT) [16]. The SPRT was derived in turn from the Neyman and Pearson [17] most powerful test for a simple hypothesis.

The decision rule of the CUSUM chart runs off the cumulative statistic

$$C_t = \sum_{j=1}^{t} (Z_i - k)$$

where $Z_i$ is the standardized normal of each observation,

$$Z_i = \frac{(\bar{x}_i - \mu_0)}{\sigma_{\bar{x}}}$$

and the common choice for $k$ is 0.5 [18], which corresponds to a standardized magnitude of change of 1. The CUSUM control chart sequentially compares the statistic $C_t$ against a control limit $A'$ until $C_t > A'$. Since we are not interested in concluding that the network is unchanged, the cumulative statistic is

$$C_t^+ = \max\{0, Z_t - k + C_{t-1}^+\}$$

The statistic $C_t^+$ is compared to the constant control limit, $h^+$. If $C_t^+ > h^+$, then the control chart signals that an increase in a network measure has occurred. Since this rule only detects increases in the mean, a second cumulative statistic rule must be used to detect decreases in the mean.

$$C_t^- = \max\{0, -Z_t - k + C_{t-1}^-\}$$

which signals a decrease in a network measure's mean when $C_t^- > h^-$.

The CUSUM control chart was selected for two reasons. First, this chart is well suited to detecting small changes in the mean of a process over time. In terms of a semantic content, this is a desired quality because we are interested in detecting subtle change in semantic content over time, alerting management to potential issues. By casual observation, one could conclude that a person's typical conversations generally stay the same from day to day and not expect drastic changes. Conversely, drastic changes in semantic content can be quite obvious. Since the CUSUM is good at detecting slight changes, it may be able to provide early warning for drastic changes, or reveal when more subtle changes have occurred. A second benefit of the CUSUM control chart is its built-in change point detection. After the control chart signals, the most likely change point is found by tracing the C statistic back to the last time it was zero. This allows the time of the change in the network to be calculated quickly and easily. This is extremely important for our proposed methodology for semantic change detection.

The newly proposed method for detecting semantic change is demonstrated on the corpus of Enron email, which comprises 50,000 email text documents. This data set spans a sufficient time period to be meaningful (created over a period of four years, 1998-2002) and contains at least one known major organizational change point (in this case, turnover of the CEO & Chairman). The data set forms a closed network, which only includes emails between Enron employees and excludes emails either from or to individuals outside the Enron organization.

We demonstrate a method of investigating potential changes through LSA by transforming the data into a matrix of term by significant time-periods, where key terms can be identified. These key terms show the major activities related to that domain (i.e., political ties) as identified by someone other than the authors.

## III. Method

This research explores characteristics of temporal change in text classification of large data sets. We study the following characteristics: The point when a change has been detected; the magnitude of the change; and the most likely estimate of when the change originally occurred. We then map these changes to some of the organization's historical events.

Below is an outline of the approach pursued in this study:

### *A. Conducted TF-IDF on weekly email documents*

We applied the TF-IDF algorithm to weekly email documents of Enron. The TF-IDF is a measure of sentiment in documents. Character strings are given a high score, when they occur frequently in a document, yet infrequently across multiple documents. The formula for TF-IDF is given by

$$tfidf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \ln \frac{|D|}{\left|\{d_j : t_i \in d_j\}\right|}$$

where, $n_{i,j}$ is the number of character strings, $t_i$, in document $d_j$; $n_{k,j}$ is the number of total terms in document $d_j$; and $|D|$ is the total number of documents in the corpus. For a comprehensive explanation of TF-IDF refer to Sparck-Jones [19]. Because the objective of this research is to sequentially observe weekly email and make a conclusion about whether a change may have occurred, it is not feasible to conduct TF-IDF over all possible documents to include those that will occur in the future. Therefore, TF-IDF is calculated over one week's worth of documents (emails). The TF-IDF vectors for each document are then averaged to create a TF-IDF vector that represents the sentiment of that particular week.

### *B. Performed Latent Semantic Analysis (LSA)*

LSA transforms a term by document matrix into matrices of lower dimensions through singular valued decomposition (SVD). The reduced dimension is considered to be term concepts. Similar terms across documents in the term by document matrix are grouped together in concepts through the SVD process. Given that $X$ is a term by document matrix of dimensions $t$ x $d$, then the SVD is, $X = U \bullet \sum \bullet V^T$.

The matrix $U$ is the term by concept matrix, with dimension $t$ x $c$. The matrix $V$ is the concept by document matrix (concept by week in our application) with dimension $c$ x $d$. The matrix $\Sigma$ is a diagonal matrix with the $c$ singular values of $X$. The full SVD of matrix $X$ will limit the dimension $c$ to the minimum of $t$ and $d$. However, a smaller value of $c$ can be chosen to further reduce the dimensionality of $X$. For a comprehensive explanation of LSA, refer to Deerwester [20].

### *C. Performed cosine similarity between vectors of each week's concepts*

While comparing the difference between corresponding components would be difficult for generating test statistics, comparing the angle between each vector and a reference vector will allow differences to be detected. These angles can be monitored over time for change detection using statistical process control. The cosine similarity between weekly vectors of influential terms is calculated to quantify different potential measures of weekly change.

A reference vector is determined by averaging available weekly LSA vectors in the matrix V across rows. This average vector has no significance other than serving as a reference point for calculating differences between weekly vectors. If this were not considered, small trend changes in weekly sentiment would go undetected. Finding the angle between two vectors, Vi and Vj corresponding to weeks i and j respectively, is given by,

$$\theta_{i,j} = \arccos\left(\frac{V_i \bullet V_j}{|V_i||V_j|}\right)$$

, where these angles between vectors, represents change in weekly email sentiment.

### *D. Apply CUSUM control chart statistic*

Statistical process control charts [16,18] are used to detect changes in temporal data. We use the cumulative sum (CUSUM) control chart statistic to identify potential changes in the semantic content of Enron communication. The CUSUM has the additional feature of providing an estimate of when the measure changed, in addition to signaling that it changed.

With control charts helping to distinguish process abnormality, measurements from the process are recorded then used to compute a test statistic. When the test statistic exceeds the limits of the control chart, the process is deemed abnormal. This indicates that a change in the process may have occurred. The process (in this case group sentiment) can then be investigated to identify the potential cause of the change. The CUSUM statistic is given by,

$$C_x^+ = \max\left\{0, \frac{\theta_{x,\bar{x}} - \bar{\theta}}{s_\theta} - \frac{\delta}{2} + C_{x-1}^+\right\}$$

where $\bar{\theta}$ is the average cosine similarity between a weekly vector and the reference vector when the sentiment is not

changing; and δ is the magnitude of change that the CUSUM is optimized to detect. For this study, δ = 1 for all calculations. A control limit of 2.03 was used, which corresponds to a type I error of 0.05. This allowed the weekly data to be broken into time periods of similar sentiment.

### *E. Averaged Concept vectors between change points*

We have now determined multiple potential change points in the temporality of the data. In order to identify the potential causes of change or differences between time periods, it is of interest to identify what concepts correspond to significant time periods. We therefore averaged the concept vectors over the weeks between potential change points to create a concept by time-period matrix, *G*. This matrix is then substituted for *V* in the SVD of *X*. The matrix *X** is then the term by time-period matrix, given by, $X^* = U \bullet \sum \bullet G^T$.

This new representation of the data associates salient character strings with significant time-periods.

### *F. Identified most salient character strings for each time period*

We then identified the most salient character-strings for each time period by ranking the terms according to their transformed TF-IDF scores for each time period. In addition, the biggest changes in salient terms between time-periods were calculated by taking the absolute value of the difference between sequential time-period vectors.

### *G. Matched character strings against Enron history*

We manually compared the character strings to those in Enron's history. We present those with the largest changes.

## IV. Results

The newly proposed method of change detection is demonstrated on the Enron email corpus. First significant time periods are identified. Then salient character strings for each time period are identified. Qualitative analysis is used to support the quantitative findings of the newly proposed approach.

There were 11 time-periods identified in the Enron data. The time periods are shown in Table 1.

Table 1. Significant Change Points in *θ*.

| Period | From | To |
|---|---|---|
| 1 | 13 Feb 2000 | 25 Jun 2000 |
| 2 | 25 Jun 2000 | 1 Oct 2000 |
| 3 | 1 Oct 2000 | 17 Dec 2000 |
| 4 | 17 Dec 2000 | 25 Feb 2001 |
| 5 | 25 Feb 2001 | 1 Apr 2001 |
| 6 | 1 Apr 2001 | 10 Jun 2001 |
| 7 | 10 Jun 2001 | 12 Aug 2001 |
| 8 | 12 Aug 2001 | 16 Sep 2001 |
| 9 | 16 Sep 2001 | 20 Jan 2002 |
| 10 | 20 Jan 2002 | 5 May 2002 |
| 11 | 5 May 2002 | 25 Jun 2002 |

Below are the salient character strings identified for the time-period vectors in matrix *G*:

> Dec 01; Columbia; Bridgeline; Meters; 3Jan; 16 Jan; 19 Jan; 22 Feb; 21 Feb; 12 Nov; Reallocation; Oil; California; Davis; Electricity; Commission; Govenor; Utility; Unify; Sanctions; gang; ectcc.

Sequentially taking the difference of the time-period vectors above provide the largest changes in salient character strings between time-periods:

> Energy; gas; Bridgeline; Enron; 14,15,16,17,18,19; 55,56; Oil; Meters; Energy; Trading; July; Dec 01; Columbia; 23-24; State; California; power; Davis; Gas.

At first there did not seem to be much information in the character strings above. We began by looking at significant events that occurred on the dates indicated in the character strings above. On 20 January 2001, the CEO of Enron, Kenneth Lay, and the president of Enron, Jeffery Skilling, attended the Presidential Inauguration of George W. Bush, and both donated $100,000 to the event. The character strings "16 Jan" and "19 Jan" lead up to this event. Lay and other Enron officials met with Vice President Dick Cheney energy task force on 22 February, which corresponds to the character strings "22 Feb" and "21 Feb". The string "Energy" is identified for the time-period of the California energy crisis (2000), the period that Dick Cheney's Energy Task Force is formed (Jan-Mar 2001), and the energy policy discussions with Dick Cheney in October 2001. Since much of the salient character strings surround Vice President Dick Cheney, the TF-IDF scores related to the president and vice president of the United States from matrix *X* were examined over each week and are displayed in Figure [1]. The Presidential Inauguration occurred at the end of Week 65. It can therefore be seen that the character strings "Bush" and "Cheney" show high TF-IDF scores during time leading up to the presidential election and continuing up to the Inauguration. It is not clear what connection there was between Enron and the White House after Week 120 (February 2002).

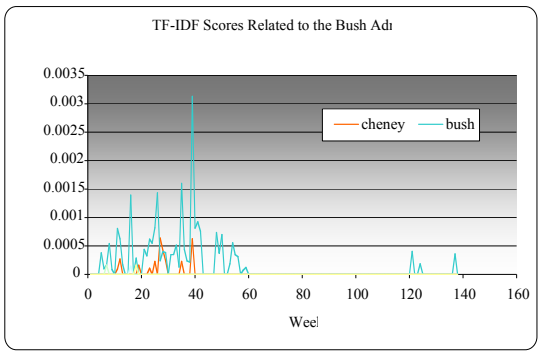


Figure [1]

There were only two major historical events involving the Vice President and Enron that were missed in this process. One is the 7 August 2001 meeting between Dick Cheney and a German subsidiary of Enron, however, since this is a European company, email traffic of this event is likely absent from the data set. The 17 April 2001 meeting between Dick Cheney and Enron officials is also missed. This meeting coincides with the Fortune Magazine Article that questions Enron's stock price; legal questions raised about LJM, a company used to hide Enron debt; and problems with the Raptor partnership. Enron repurchased Chewco's investment in JEDI to cover the problem. It is not unreasonable that these events overwhelm the significance of a meeting at the White House. During this time salient terms are "reallocation" and "trading" which further support the likelihood that Enron was facing more important issues than the Vice President.

The salient terms California, Davis, Governor, and state occur in the time-period 10 June – 12 August 2001. This corresponds to CEO of Enron, Jeffery Skilling, joking about the California energy crisis at a Las Vegas conference, and the subsequent media backlash during his visit to California later in the month. Figure [2] shows the TF-IDF scores for "California" and "Davis" for each week.

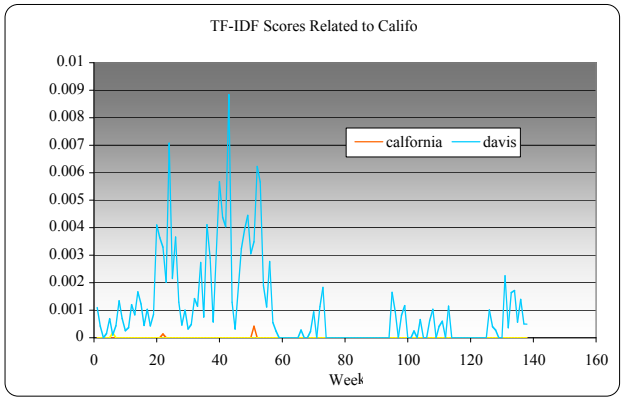


Figure [2]

The large activity in the early weeks of Figure [2] correspond to the California energy crisis. The activity following week 90 can be attributed to Skilling's joke. There is insufficient historical information on Enron to explain further activity in the chart.

During the period September 2001-January 2002, the character string "@ectec" became salient. We found that this is a secure email service. This is also the period that Enron was accused of being an "elaborate accounting hoax". Arthur Andersen Consulting Firm hires a law firm to prepare a defense, and essentially all of the Enron legal issues surface to the public. It appears some employees may have begun using an external secure email service at this time.

Several other salient terms to include "Oil", "Electricity", "Commission", "Utility", "Sanctions", "Gas", "Meters", and "Power" were not associated with any significant historical events. Considering that Enron was an energy company, these terms are consistent will normal business practices. There were four character strings that had no apparent explanation: "Dec 01", "Columbia", "Bridgeline", and "gang". This of course does not mean that there is not some significant event associated with these character strings that could not be found in the literature.

We have shown a method to reduce the immense data of over 50,000 emails to a manageable size through LSA, and then apply statistical process control to detect possible changes in a social group for investigation. We demonstrated a method of investigating potential changes through LSA by transforming the data into a term by significant time-period matrix, where key terms could be identified. These key terms show the major activities related to that domain (political ties) as identified by someone other than the principal investigators.

## V. Conclusion

We have demonstrated a novel method of detecting change in semantic content of text over time. Latent semantic analysis reduces the dimensionality of text documents. Cosine similarity provides a measure of similarity between time periods in the reduced data set. Statistical process control can then be applied to detect meaningful change. The latent semantic analysis can then be reversed to identify a term by significant time period matrix that can be used to search for the cause of change. This last step is perhaps the most novel innovation of this paper. Its' effectiveness was qualitatively demonstrated on the Enron email corpus.

The application in public policy is clear. For maximum impact, a policy maker's representation of events must reflect constituent beliefs. Placing objective, quantitative measures on these beliefs can make for more responsive, if not better, governance.

Also benefiting from this approach is the task of product or service commercialization. Customer adoption can be powerfully impacted by word-of-mouth. The targeting of customer referrals could be made more effective, and addressing customer complaints could occur before becoming a material issue.

Risk analysis in finance can benefit from all studies into detecting change. Issues as diverse as securities or currency speculation, unstable government policy, tax-avoidance schemes, accounting policy changes, or changes in credit standards are reflected in Natural Language in addition to numbers. Detecting changes in Natural Language may be an important adjunct to changes in the numerical data. Increasingly organizations examine email to protect themselves against corporate malfeasance [6].

Yet another application of this approach is in the domain of security and law enforcement. To the degree that unlawful activities are reflected in sentiment, automated processes for detecting changes are inherently more effective than manual processes in the volume of data able to be processed.

The sheer scale of the text to be classified in the Enron email dataset gives it substantial interest to researchers. While there has been study on corpra of substantial size, few, if any, have been allowed at this scale on email. Analysis can be done from web log (or 'blog') communication with a virtually unlimited data set. However, email can represent a more active dialogue with communications occurring more rapidly with the subject changing quickly and in discontinuous spurts, with the addition of the time stamp adding to the data's richness. The very large data set that we have investigated has been the subject of a great deal of interest but remains rich enough to justify continued study. Our collection of 50,000 emails actually represents a subset of a larger data set where at least 250,000 emails are available. Should the larger set become usable, it will be of interest to at least these researchers. Unfortunately, this larger set suffers difficulties requiring further cleaning (e.g., unreconciled to/from fields)

There exist at least three straightforward extensions to the work presented in this paper including the application of additional classification algorithms to the existing Enron corpus investigated here, applying the classification approaches to other data sets, and applying the methodology for the purpose of predicting changes/events for the social network described by the email communications. As a scientific community we can hope to see more research in this area as classification algorithms and methods of statistical process control continue to improve.

## ACKNOWLEDGMENT

We are grateful to Geoff Gordon, and Ziv Bar-Joseph for their feedback. Jana Diesner and Terrill Franz provided assistance in the preparation of the original dataset. The authors also wish to thank Peter Landwehr and Brian Hirshman for their helpful comments making the study's conclusions more clear.

## REFERENCES

[1] Lawrence, S., & Giles, L. (1999). Accessibility and Distribution of Information on the Web. *Nature, 400*, 107-109.

[2] UCBerkeleyLibrary. (2007). *The BEST Search Engines*. Retrieved November, 2007, 2007, from http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/SearchEngines.html

[3] Mani, I., & Bloedorn, E. (1997). Summarizing Similarities and Differences Among Related Documents. *Information Retrieval*(1), 35-67.

[4] Stockinger, K., Rotem, D., Shoshani, A., & Wu, K. (2006). *Analyzing Enron Data: Bitmap Indexing Outperforms MySQL Queries by Several Orders of Magnitude* [Electronic Version], 4. Retrieved 2006 Jan 28.

[5] Carley, K. M., & Skillicorn, D. (2005). Special Issue on Analyzing Large Scale Networks: The Enron Corpus. *Computational & Mathematical Organizational Theory* (11), 179-181.

[6] Keila, P. S., & Skillicorn, D. B. (2005). Structure in the Enron Email Dataset. *Computational & Mathematical Organization Theory* (11), 183-199.

[7] Priebe, C. E., Conroy, J. M., Marchette, D. J., & Park, Y. (2005). Scan Statistics on Enron Graphs. *Computational & Mathematical Organization Theory* (11), 229-247.

[8] Godbole, N., Srinivasaiah, M., & Skiena, S. (2007). *LargeScale Sentiment Analysis for News and Blogs (System Demonstration)*. Paper presented at the International Conference for Weblogs and Social Media (ICWSM 07), Boulder, CO.

[9] Li, Y. H., & Jain, A. K. (1998). Classification of Text Documents. *The Computer Journal, 41*(8), 537-546.

[10] Hynek, J., & Jezek, K. (2003, 25-28 June 2003). *Practical Approach to Automatic Text Summarization*. Paper presented at the 7th ICCC/IFIP International Conference on Electronic Publishing, Universidade do Minho, Portugal.

[11] Chuang, W. T., Tiyyagura, A., Yang, J., & Giuffrida, G. (2000). *A fast algorithm for hierarchical text classification*. In Y. Kambayashi, M. Mohania & A. M. Tjoa (Eds.), *DaWaK 2000* (Vol. LNCS 1874, pp. 409-418): Springer-Verlag Berlin Heidelberg 2000.

[12] Aizawa, A. (2000). *The feature quantity: an information theoretic perspective of Tfidf-like measures.* Paper presented at the Annual ACM conference on research and development in information retrieval, Athens, Greece.

[13] Jing, L.-P., Huang, H.-K., & Shi, H.-B. (2003). *Improved feature selection approach TFIDF in text mining*. Paper presented at the International Conference on Machine Learning and Cybernetics, 2002. Proceedings. 2002

[14] Moon, N., & Singh, R. (2005). *Experiments in Text-Based Mining and Analysis of Biological Information from MEDLINE on Functionally-Related Genes* Paper presented at the 18th International Conference on Systems Engineering (ICSEng'05).

[15] Ishii, N., Murai, T., Yamada, T., & Bao, Y. (2006). *Text Classification by Combining Grouping, LSA and kNN.* Paper presented at the 5th IEEE/ACIS International Conference on Computer and Information Science and 1st IEEE/ACIS International Workshop on Component-Based Software Engineering, Software Architecture and Reuse (ICIS-COMSAR'06).

[16] Page, E. S. (1961). Cumulative Sum Control Charts. *Technometrics, 3*(1), 1-9.

[17] Neyman, J. & Pearson, E.S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Phil. Trans. Royal Society Series A.* 231, 289-337.

[18] Montgomery, D. C. (1996). *Introduction to Statistical Quality Control* (3rd edition): John Wiley & Sons.

[19] Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28 (1), 11-21.

[20] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1988). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391-407.