

Entropy Based Measure Functions for Analyzing Time Stamped Documents

Parvathi Chundi Rui Zhang
C.S Dept., University of Nebraska at Omaha
Omaha, NE 68182
{pchundi, rzhang}@mail.unomaha.edu

Malu Castellanos
HP Labs
Palo Alto, CA 94304
malu_castellanos@hp.com

Abstract

Measure functions that assign numeric values to keywords to capture their significance in a document set play a crucial role in the construction of a time decomposition of a document set. In this paper, we define two measure functions based on the notion of entropy. The interval entropy measure function identifies time intervals that have non-uniform keyword distributions and assigns high measure function values to keywords with high relative occurrence frequency in that time interval. The keyword entropy measure function similarly identifies keywords that have non-uniform occurrence frequency over time. The measure functions are applied to construct several time decompositions of a subset of documents from the TDT-Pilot corpus and the Enron Email data set. The results indicate that the measure functions are highly effective in capturing the temporal content of the document set.

Keywords: entropy, time decomposition, measure function, time stamped document sets, information loss, information content

1 Introduction

Text documents have become widely accessible virtually in all types of domains. Almost all of the available documents contain some sort of time stamp, such as publication date, indicating the time the information in the document was compiled. The publication time stamp can be used to map the document data into the temporal dimension by simply assuming that all topics/keywords occurring in the document set occur during the publication date of the document. By assigning a temporal dimension to document data, one can discover temporal trends, correlations, and other temporal information from the document set.

Extracting temporal information from time stamped document sets has been an active area of research [1, 3, 4, 5, 6, 8, 9, 10, 11]. A common method for explicating the temporal information hidden in a document set is to construct a *time decomposition* of

the document set. The time period associated with the document set is partitioned into one or more time intervals. Each document in the document set is assigned to exactly one of the time intervals based on its time stamp. A numeric measure is then assigned to each of the keywords appearing in the document set belonging to a time interval using functions such as count and ratio. The information of each time interval is then represented as a bag of keywords that are determined to be significant based on the assigned measure. The sequence of keyword sets, one for each time interval in the time decomposition, can then be analyzed for temporal trends and other temporal information.

In our earlier work [3, 4, 5, 6], we studied the problem of constructing optimal information preserving and information lossy time decompositions of a document set. We illustrated the effectiveness of time decompositions for capturing temporal content of a document set through several experimental results.

Functions, which we call *measure functions*, that calculate a measure of significance of keywords play an important role in explicating the temporal information of the document set. In general, there are two notions of temporal information – information pertaining to specific time intervals and information regarding how document data changes over time. As an example, consider news coverage. Days where there are a few news stories that get a lot more coverage (E.g. bird flu outbreaks) than other stories are a lot more interesting than days where all stories get the same kind of coverage. Similarly, if a news item gets a similar amount of coverage over a long period of time, people tend to lose interest in the story. This happens even if that story happens to receive a lot of coverage during that time period. On the other hand, if the amount of coverage shows temporal fluctuations, then it becomes worthy of attention.

The topic of measure functions received very little attention so far in the area of temporal text mining. Most of the previous work including our earlier work [3,

4, 5, 6] employed simple functions such as count measure and ratio measures to determine the significance of a keyword in a time interval. The above measures coupled with a threshold mechanism are good at identifying keywords that have high absolute frequencies in a time interval or have a high ratio of occurrences. However, they fail to capture information in time intervals where the absolute values of frequencies or ratios may not be high. Consider two intervals. In the first time interval, topic *A* occurs 3 times, topics *B*, *C*, ..., *G* appear 1 time each. In the second time interval, topics *A*, *B* and *C* each appear 10 times. A threshold of 0.33 will make topic *A* significant on both cases. However, the occurrence of *A* is more significant in the former time interval due its high relative frequency whereas in the latter time interval all topics receive equal attention.

Count and ratio measure functions do not take into account the frequencies of keywords on other time intervals. A topic such as *Iraq war* may receive high attention consistently and hence is deemed significant by the above measures. However, it carries little temporal information when compared to topics that have occasional spikes in the coverage such as the *anthrax investigation*. Also, one may be interested in time periods where the topic may have received more coverage than the rest of the time intervals. This information is not directly measured by count and ratio measures.

In this paper, we introduce two measure functions based on simple information theoretic notions [7]. Based on our knowledge, this is perhaps the first time information theoretic notions such as entropy have been used to extract temporal information from time stamped document sets. Entropy function measures the uncertainty associated with a given distribution of probabilities. Uniform probability distributions such as days with no headlines will have a higher entropy value than days with headlines. Similarly, keywords with temporal fluctuations in frequency of occurrence will have a lower entropy than keywords with consistently high (or low) occurrence over time.

The **interval entropy** measure function identifies keywords that contribute to lowering the entropy of the keyword distribution of a time interval. Let P be the probability distribution of a time interval. Consider a keyword w . For w to be significant in the time interval, the entropy of P in the absence of w must be higher than the entropy of P including w . This will be true if a keyword has a higher relative frequency than other keywords in the time interval. The higher the relative frequency of w in P , the higher the change in entropy. This fact can be used in determining a value for the threshold α .

The second measure function **keyword entropy** measures the entropy of the distribution of a keyword over time. Consider the frequency distribution of a keyword in each time interval of a time decomposition of the document set. A keyword is considered significant in a time interval if its frequency of occurrence in that time interval is relatively high compared to other time intervals. We use the above notion of change in entropy to identify keywords that have non-uniform distributions of occurrences over time.

To study the behavior of these measure functions, we applied them to extract the temporal information from a subset of Reuters news articles published during the months of July and August 1994 and a subset of Enron email messages. We constructed several time decompositions of each document set using the two measure functions and a simple ratio function. We studied the temporal information captured in these time decompositions using the notions of *information content* and *information loss* [3, 4, 5]. The results are summarized below.

1. All three measures were good at capturing highly frequent keywords in a time interval. However, interval entropy measure was better at capturing low frequency keywords in some intervals. And, keyword entropy measure identified rarely occurring keywords.
2. The interval entropy measure function was very good in identifying keywords in each interval whose frequency of occurrence was high relative to other keywords in the same interval regardless of the absolute values of occurrence frequencies. The behavior of the interval entropy measure is observed to be similar to a *stable* measure function [3]. We have yet to prove the claim.
3. The keyword entropy measure successfully identified time intervals in which a keyword has a relatively high frequency. However, the function also characterized lots of keywords that appear just once (or very few times) in just one interval as significant. Such keywords satisfy the definition of the keyword entropy measure. As a consequence, the number of keywords deemed significant by the keyword entropy measure tended to be large. These keywords can be removed in the cleanup stage of the process if it can be determined that they are not useful at the application level.
4. The information loss value fell as the number of intervals permitted in a time decomposition increased. This was true across for three measure functions tried.

The rest of the paper is organized as follows. Section 2 discusses some preliminaries. Section 3 defines the measure functions and illustrates them with examples. Section 4 presents the experimental results. In Section 5, we discuss some related work. Section 6 discusses conclusions and future work.

2 Preliminaries

A *time point* is an instance of time with a given *base granularity*, such as a second, minute, day, month, year, etc. A *time interval* is a sequence of one or more consecutive time points. The **length** of a time interval T , denoted $|T|$, is the number of time points within T . We use $T_{x,y}$ to denote a subinterval that includes from the x^{th} time point through the y^{th} time point of the time period.

A **decomposition** Π of a time interval T , is a sequence of subintervals T_1, T_2, \dots, T_k , such that T_{i+1} immediately follows T_i for $1 \leq i < k$, and T equals the concatenation of the k time intervals, which we write as $T = T_1 * T_2 * \dots * T_k$. Each T_i is called a **subinterval** of Π . The **size** of decomposition Π is the number of subintervals k in Π . The time interval associated with decomposition Π is denoted as $T(\Pi)$. The **shortest interval decomposition** Π_S of a time interval T is the decomposition with $|T|$ subintervals, one for each time point within T . Each subinterval within Π_S is called a **base interval**.

A decomposition Π_U of a time interval is a **uniform length decomposition** if each subinterval in Π_U contains the same number of time points. For example, the shortest interval decomposition is a uniform decomposition where each interval contains a single time point.

We now describe the relationship between time stamped documents and time points, intervals and decompositions. Consider a finite set of documents D where each document has a *time stamp* denoting its time of creation or publication. To map these documents to the time domain, we identify a time stamp in a document with a time point. (This implies that time stamps in all documents in D have the same base granularity. If not, or if the time stamps are too fine-grained, we assume that all time points can be converted to an appropriate base granularity.) Given a decomposition, each document is assigned to the subinterval in the decomposition that contains its time stamp.

Given a keyword w and a document set D , a **measure function** f_m assigns a value to keyword w that denotes a measure of significance of w in D . We assume that this value is a nonnegative real number. We also assume that if w does not appear in D , then $f_m(w, D) = 0$. We refer to v as a **measure function value** or as a **sig-**

nificance value. Depending on the characteristics of a given measure function f_m , a keyword w may need to have a high measure function value (at or above a specified threshold) or a low measure function value (at or below a specified threshold) to be significant.

The **information content** of a document set D for a given measure function f_m and a threshold $\alpha \in R^+$ is the set of keywords w appearing in D such that $f_m(w, D) \geq \alpha$ (or in some cases at most α). The *information content* of a time interval T , denoted as $I_\alpha(T, f_m)$, is the information content of the document set assigned to it. The **information content** of a decomposition $\Pi = T_1 * \dots * T_k$, denoted as $I_\alpha(\Pi, f_m)$, is $\bigcup_{i=1}^k I_\alpha(T_i, f_m)$.

Note that $I_\alpha(\Pi, f_m)$ is not necessarily equal to $I_\alpha(T(\Pi), f_m)$. ($T(\Pi)$ is the time interval associated with the decomposition Π .) In fact, the information content of different decompositions of the same document set may be different, both in terms of the cardinality and contents of the keyword set [5].

To compare different decompositions of a document set, a measure based on loss of information was introduced in [3]. Given a time interval T_i , let $T_{i1} * T_{i2} * \dots * T_{iq}$ be the time points in T_i . We define the **information loss** (μ_j) between the information contents of T_i and a time point T_{ij} ($1 \leq j \leq q$) to be the size of the symmetric difference between $I_\alpha(T_i, f_m)$ and $I_\alpha(T_{ij}, f_m)$. Then, the information loss of T_i , denoted by $\mu(T_i)$, is defined to be $\sum_{j=1}^q \mu_j$. The information loss of a decomposition is the sum of information losses for each of its subintervals. A decomposition $\Pi(T)$ of a time interval T is **lossy** if its information loss is nonzero.

3 Entropy Based Measure Functions

A measure function plays an important role in identifying which keywords/topics are significant in a given document set and consequently in a time interval. Different measure functions may lead to different sets of significant keywords [5]. In our earlier work, we defined simple count-based measure functions which deem frequently occurring (or frequency of occurrence above a certain threshold) keywords as significant.

In this paper, we employ the notions of entropy to extract temporal information from the document set [7]. Let $W = w_1, \dots, w_k$ be a set of keywords from a document set D . Let X be a random variable. The probability that X takes on value w_i is simply the ratio of the number of occurrences of w_i in D to the total number of occurrences of all keywords w_1, \dots, w_k in D . The probability distribution over keywords in D can be used to decide if the document set is interesting. If the probability distribution is uniform, then all keywords are equally likely in D and hence, the distribution is

not interesting.

Let $P = p_1, \dots, p_k$ be a probability distribution. The entropy of P , $H(P) = -\sum_{1 \leq i \leq k} p_i \log_2(p_i)$. A uniform distribution of probabilities has the highest entropy. For example, if we have two keywords with probability 0.5 each, entropy of such a distribution is a 1. Any other distribution of probabilities over the two keywords would have an entropy less than 1. Therefore, entropy value less than 1 represents the scenario where one of the keywords is more prevalent than the other.

We use the notion of change in entropy of a probability distribution over a document set to identify keywords that are significant in that document set.

The idea behind the measure functions is simple. Let W be the set of keywords of a document set D and let $W' = W - \{w\}$. Suppose we wish to find if a keyword $w \in W$ is significant in D . The effect of the presence of w on the rest of the keywords in W' is computed by the change in the contribution of the keywords in W' to the entropy of D in the presence and absence of w . If w occurs a lot more than the keywords in W' , the contribution of the keywords in W' to the entropy of D will be smaller in the presence of w . This is because for any keyword $w_j \in W'$, probability of w_j , $pr(w_j)$ will be small and consequently $pr(w_j) \log_2(pr(w_j))$ will be small. If w occurs just as frequently as keywords in W' , $pr(w_j)$ value may be (almost) same whether or not w is present.

Let $tot(W)$ denote the total number of occurrences in D of all keywords in W and $freq(w_i)$ denotes the occurrence frequency of a single keyword w_i in D . The contribution of all keywords in W' to the entropy of the document set when $w \in W$, denoted by $H(W')$ is shown below.

$$H(W') = -\sum_{w_j \in W'} pr(w_j) \log_2(pr(w_j)).$$

The probability of a w_j in W' in the absence of w , denoted by $pr_{-w}(w_j)$ is $freq(w_j)/tot(W')$. The contribution of keywords in W' to the entropy of D in the absence of w , denoted by $H_{-w}(W')$ is shown below.

$$H_{-w}(W') = -\sum_{w_j \in W'} pr_{-w}(w_j) \log_2(pr_{-w}(w_j)).$$

The difference between $H_{-w}(W')$ and $H(W')$ can be used to determine if w has a relatively high occurrence frequency in the document set D . If the difference is small, then w is not significant. On the other hand, if the difference is large, w 's occurrence is high when compared to other keywords in D .

We define the **interval entropy** (f_{ie}) measure function as $f_{ie}(w, D) = H_{-w}(W') - H(W')$. The following examples illustrate the working of f_{ie} .

Example. Suppose a, b , and c each occur 10 times in D .

To compute $f_{ie}(a, D)$, we need to compute $H_{-a}(\{b, c\})$ and $H(\{b, c\})$. Here $W' = \{b, c\}$.

In the absence of a , probability of b and c are both 0.5. Then, $H_{-a}(\{b, c\}) = -\sum_{b, c} p_i \log_2(p_i)$ which is 1. In the presence of a , the probabilities of b and c are both 0.33 and $H(\{b, c\}) = 1.055$. Hence, $f_{ie}(a, D) = -0.055$.
 $f_{ie}(b, D) = f_{ie}(c, D) = -0.055$.

In the above example, the contribution of keywords in W' to the entropy of D is almost the same in the absence of a (or b or c) and in the presence of a (or b or c). This indicates that none of the keywords have high relative frequencies.

Example. Let $W(D) = \{a, b, c\}$. Let the frequency of occurrence of keywords be as follows: a occurs 10 times, b occurs 5 times, and c occurs 3 times.

$$f_{ie}(a, D) = H_{-a}(\{b, c\}) - H(\{b, c\}).$$

$H_{-a}(\{b, c\}) = -(5/8) \log_2(5/8) - (3/8) \log_2(3/8)$ which is $0.424 + 0.531 = 0.955$.

$H(\{b, c\})$ is calculated as follows. $pr(b) = 5/18$, and $pr(c) = 3/18$ and is $0.51 + 0.43$ and is 0.94 .

Therefore, $f_{ie}(a, D) = H_{-a}(\{b, c\}) - H(\{b, c\}) = 0.01$.

$f_{ie}(b, D)$ and $f_{ie}(c, D)$ can be similarly calculated as -0.12 and -0.06 .

The contribution of $\{b, c\}$ to the entropy of the above keyword distribution is higher if a is not considered. If the occurrences of a is also added, then the entropy falls. As the frequency of a increases as compared to others, $H(\{b, c\})$ decreases further causing larger differences in entropies. For example, suppose the frequency of a is 100. Then, $H(\{b, c\}) = 0.348$ and $f_{ie}(a, D) = 0.607$.

The information content of D is the set of all keywords w whose measure function values are above a user specified threshold α . If $f_{ie}(w, D) < 0$, then the contribution of W' to the entropy of D increases when w is added and hence w is not significant in D . If $f_{ie}(w, D) \geq 0$, then the amount of positive change can be used as a threshold. A formula for deciding a value for α based on properties of H is still under development.

Therefore, given α , $I_\alpha(D, f_{ie}) = \{w | f_{ie}(w, D) \geq \alpha\}$. Note that for $\alpha = 0$, $I_\alpha(D, f_{ie})$ in Example 1 is a null set whereas $I_\alpha(D, f_{ie})$ in Example 2 is $\{a\}$.

The second measure function, which we call the **keyword entropy** considers the distribution of a keyword over all time points or time intervals. If a keyword occurs uniformly in all time points/intervals, it contributes little to the temporal aspect of the document set. If it occurs frequently at one time or another, it is considered interesting.

Let $\Pi = T_1 * T_2 * \dots * T_n$ be a decomposition of the time period associated with the given document set D . A time decomposition of D partitions D into subsets

$Docs(T_1), \dots, Docs(T_n)$. Let us suppose the occurrence frequency of w in $Docs(T_i)$ is $f_i(w)$. Let $tot(w)$ denote the total $\sum_{1 \leq i \leq n} f_i(w)$. Then, the probability of occurrence of w in T_i , $pr_i(w)$, is $f_i(w)/tot(w)$. The entropy of the probability distribution of w , $pr_1(w), \dots, pr_n(w)$ can be used to compute a measure function value for w in a document set $Docs(T_i)$.

The keyword entropy measure function, f_{kw} , also computes two entropy values $H(D')$ and $H(D - Docs(T_i))$ and assigns the difference between these two values as a measure function value to w in $Docs(T_i)$. $H(D')$ is the entropy of the probability distribution $freq_1(w)/tot(w) - freq_i(w), \dots, freq_{i-1}(w)/tot(w) - freq_i(w), freq_{i+1}(w)/tot(w) - freq_i(w), \dots, freq_n(w)/tot(w) - freq_i(w)$. $H(D - Docs(T_i))$ is the entropy of the distribution $freq_1(w)/tot(w), \dots, freq_{i-1}(w)/tot(w), freq_{i+1}(w)/tot(w), freq_n(w)/tot(w)$. The working of f_{kw} is illustrated by the example below.

Example. Keyword a appears 10 times in $Docs(T_1)$, 5 times in $Docs(T_2)$ and 3 times in $Docs(T_3)$. Keyword b appears 3 times each in $Docs(T_1)$, $Docs(T_2)$ and $Docs(T_3)$.

To decide if a is significant in $Docs(T_1)$, we compute $H(D')$ as $H(5/8, 3/8)$ and $H(D - Docs(T_1))$ as $H(5/18, 3/18)$. Therefore $f_{kw}(a, Docs(T_1)) = 0.01$. $f_{kw}(a, Docs(T_2))$ and $f_{kw}(a, Docs(T_3))$ can be similarly computed.

Similarly $f_{kw}(b, Docs(T_1)) = H(D') - H(D - Docs(T_1))$. It is $H(1/2, 1/2) - H(1/3, 1/3)$ which is -0.055 . $f_{kw}(b, Docs(T_2))$ and $f_{kw}(c, Docs(T_3))$ can be similarly computed.

As we discussed above, an α value of zero or greater can be used to identify significant keywords for each $Docs(T_i)$. Finding an appropriate value of α for a given keyword distribution is our future work.

4 Experiments

In this section, we describe results from some preliminary experiments conducted using two data sets. The first data set is a subset of *TDT-Pilot* corpus (available at <http://projects.ldc.upenn.edu/TDT-Pilot>). The *TDT-Pilot* corpus contains 16,000 stories collected from Reuters newswire and CNN broadcast news transcripts during the period from 1st July 1994 to 30th June 1995. Our experiments included only the titles from Reuters news articles from 1st July 2004 to 31st Aug 2004. Each title was treated as a separate document in the experiment. There were 1103 total documents in the document set.

The second data set contained a subset of email messages from the Enron Email data set (available at

<http://www-2.cs.cmu.edu/enron>) and contains about a million emails messages exchanged among the senior management personnel at Enron Corporation. Our analysis was conducted on approximately 1700 email messages sent during Apr 1st, 2002 — Dec 21st, 2002. Each email message was treated as a separate document and time stamped with the day it was sent.

For our experiments, the base granularity for both document sets was chosen to be one day to avoid having a sparse distribution of the data. There were 62 intervals in the shortest interval decomposition of the *TDT-Pilot* corpus. There were 84 time points in the shortest interval decomposition of the Enron data set after removing all time points with empty content.

The data sets were then prepared for analysis as follows. Each document was tokenized, and non-noun keywords and stop words, if any, were removed. We further limited our analysis to the top 10 noun keywords in each time point. For each keyword, count of how many documents (titles) include it in each time point (day) was recorded.

We applied three measure functions to documents in each time point to compute the information content. In addition to the two measure functions defined earlier, we also defined a simple ratio measure f_r as follows. $f_r(w, D)$ is the ratio of the number of occurrences of w in D to the total number of occurrences of all keywords in D .

4.1 Reuter's Data Set We illustrate the effect of measure functions using the documents published in the first week of July 1994. Table 1 lists some of the top 10 keywords for each time point in the first week of July 1994¹. The number in parenthesis indicates the frequency of the keyword in that time point. Table 2 shows the information content for each of the time points computed from the three measure functions. It can be seen that both f_r and f_{ie} include the high frequency keywords from the document set in each time point's information content. In some cases, the information content of a time point computed by f_r is null (denoted by *NULL* in the table) whereas f_{ie} identifies keywords with a relative high frequency (e.g: time point 1).

The information content computed by f_{ke} includes keywords that uniquely occur in a time point (e.g.: *stall* in time point 2, *Mukalla* in time point 4). If a keyword occurs in more than one time point, such as the keyword *US* or *Gaza*, then such a keyword appears

¹Please note that the observations extend to any arbitrary set keywords.

| No. | Time point | Keywords |
|-----|------------------------------|--|
| 1. | July 1 st , 1994 | Gaza(2), Simpson(2), knife (2), recall(2), fight(1), aid(1), storm(1), peacekeeping(1), surgeon(1), no(1) |
| 2. | July 2 nd , 1994 | Arafat(4), dead(3), palestinian(3), shot(2), evidence(1), Algeria(1), communist(1), storm(1), stall(1) |
| 3. | July 3 rd , 1994 | visit(2), french(2), Arafat(2), Gaza(1), reform(1), Citadel(1), third discriminate(1), police(1) |
| 4. | July 4 th , 1994 | troop(2), french(2), say(2), rebel(2), northern(2), Yemen(2), Muk: capture(2), evidence(1), australian(1), |
| 5. | July 5 th , 1994 | say(3), independent(2), govern(2), haitian(2), world(2), cup(2), da US(2), Christopher(1) |
| 6. | July 6 th , 1994 | world(3), right(2), cup(2), party(2), citizenship(1), south(1), homecoming(1), male(1), bank(1), start(1) |
| 7. | July 7 th , 1994. | US(3), UN(2), Serbia(1), ginseng(1), troop(1), impeach(1), peace start(1), parliament(1) |

Table 1: Keyword Frequencies

in the information content of the time point in which has the highest frequency.

Information content of time intervals with two or more time points can be similarly computed. Due to lack of space, we omit the details here.

We computed several time decompositions of the document set. The shortest interval decomposition Π_S of the document set contains one interval for each time point (62 intervals in total for the document set considered) and is a lossless decomposition of the document set. For each measure function, we computed several optimal lossy decomposition by varying the size² of a decomposition. For each size, we computed the cardinality of the information content of the time decomposition.

The plot in Figure 1 illustrates how the cardinality of the information content changes w.r.to the size of decomposition for each measure function. Here X -axis here plots the size of a decomposition and the Y -axis plots the cardinality. The cardinality of information content computed by f_{kw} is bigger than the other two measure functions. This is because the information content of an interval includes keywords even if they occur in a single time point in that interval. These keywords may not be included in information content computed by other measure functions since their relative frequency may become insignificant as intervals get larger. The cardinality of information content of Π_S is 41, 156, and 244 for f_r , f_{ie} and f_{kw} respectively. As can be seen from the figure, the information content of a decomposition increases with its size for f_r and f_{ie} . It is interesting to note that the cardinality of information content computed by f_{kw} is very close to that of Π_S even for small size decompositions. Also, the cardinality falls somewhat as the size increases. We are currently

²Size of a decomposition Π is the number of intervals contained in Π .

| No. | Fr ($\alpha = 0.15$) | $f_{ie}(\alpha = 0.2)$ | $f_{kw}(\alpha = 0.11)$ |
|-----|---------------------------|---------------------------------|---|
| 1. | NULL | Gaza, Simpson, knife, recall | Gaza, knife, recall, aid, peacekeeping, surgeon, fight, Simpson |
| 2 | palestinian, Arafat, dead | palestinian, Arafat, dead, shot | stall, dead, palestinian, Algeria, communist, shot, Arafat, dead, palestinian |
| 3 | visit, Arafat, french | visit, Arafat, french | discriminate, citadel, police, visit, reform, third |
| 4 | NULL | NULL | troop, Mukalla, northern, capture, Yemen, rebel, australian |
| 5 | NULL | Say | government, independent, face, Christopher, haitian, say, |
| 6 | world | cup, party, world, right | citizenship, homecoming, party, south, male, bank, start, world, right |
| 7 | US, UN | US, UN | Serbia, ginseng, parliament, impeach, peace, UN, US |

Table 2: Information Content

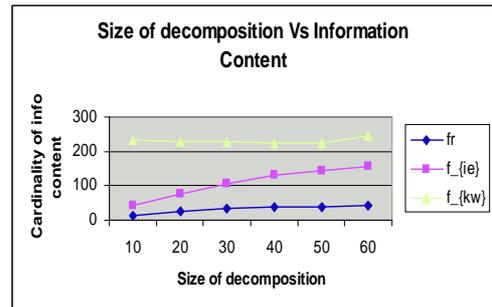


Figure 1: Size of Decomposition Vs Information Content

investigating the reason for this phenomenon.

The plot in Figure 2 shows the effect of size of a decomposition on information loss. The X -axis plots the size of the decomposition whereas Y -axis contains the information loss values. As can be seen from the figure, the information loss falls as the size of a decomposition increases. The high information loss values for f_{kw} when compared to other measure functions can be explained by the current definition of information loss.

4.2 Enron Data Set The results from the Enron data set were very similar to those from Reuters data. Tables 3 and 4 show the keyword frequencies and the

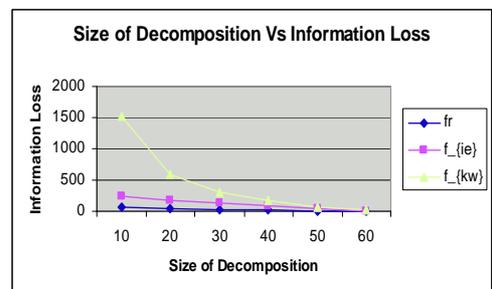


Figure 2: Size of Decomposition Vs Information Loss

| No. | Time Point | Keywords |
|-----|----------------------------|---|
| 1. | Apr 1 st , 2002 | today(5), will(5), last(4), meet(4), April(3), day(3), help(3), money(3), gas(3), request(3) |
| 2. | Apr 2 nd , 2002 | gas(5), will(5), April(4), custom(4), day(4), document(4), need(4), contract(3), entity(3) |
| 3. | Apr 3 rd , 2002 | Gas(9), activity(4), day(4), April(3), call(3), draft(3), good(3), handle(3), last(3), plant(3) |
| 4. | Apr 4 th , 2002 | will (12), data(8), gas(8), can(7), contact(7), inform(7), may(7), energy(6), call(5) |
| 5. | Apr 5 th , 2002 | will(4), call(3), deal(3), help(3), message(3), need(3), service(3), subject(3), thing(3), can(2) |
| 6. | Apr 8 th , 2002 | go(6), will(6), data(5), inform(5), last(5), romorrow(5), back(5), gas(4), number(4), service(4) |
| 7. | Apr 9 th , 2002 | can(13), will(13), day(9), group(8), may(7), April(6), don(6), look(6), request(6) |

Table 3: Keyword Frequencies from the Enron Data

| No. | $f_r(\alpha = 0.15)$ | $f_{ie}(\alpha = 0.2)$ | $f_{kw}(\alpha = 0.11)$ |
|-----|----------------------|-------------------------|--|
| 1. | NULL | meet, last, today, will | gas, money, meet, today |
| 2. | NULL | gas, will | document, need, contract, entity, custom |
| 3. | gas | gas | Activity, good, plant, handle, draft |
| 4. | will | will, gas, data | data, contact, will, inform, energy |
| 5. | NULL | will | message, subject, deal, thing |
| 6. | NULL | go, will | tomorrow, number, back, service |
| 7. | Can, will | day, can, will | request, day, look, group, don |

Table 4: Information Content for the Enron Data

information content computed by the measure functions for each time point in the first week of April 2002. The top 10 noun keywords in each time point in the Enron data set are not as informative as the keywords for the Reuters data set. This suggests that we may need to employ domain-specific methods to identify meaningful keywords from the data set. Figure 3 shows how the cardinality of information content of a decomposition changes with its size and Figure 4 plots the value of information loss of a decomposition with respect to its size. As observed with the Reuter’s data set, the information content captured by a decomposition increases with its size and therefore, the information loss decreases.

5 Related Work

Segmenting a document set based on the time stamps for identifying trends and tracking interesting topics is an active area of research. The work on topic detection and tracking in [1, 2] extracts significant topics/events from news articles by grouping the articles published on the same day together. Papers on extracting trends from time stamped text documents also use time decompositions, where subintervals are of length one year

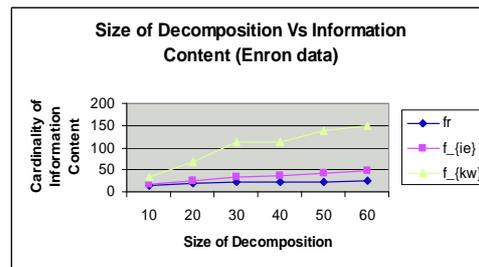


Figure 3: Size of Decomposition Vs Information Content

[10, 11] or a day [12, 13, 14]. These papers do not focus on the issue of measure functions to determine temporal significance of keywords. In [8], the author describes how to identify bursts from document streams such as news articles by modeling the streams as infinite automaton, whereas our work is more applicable to finite document sets. Our earlier work [3, 4, 5, 6] formulated the problem of constructing optimal information preserving as well as lossy time decompositions of time stamped documents and identified the crucial role played by measure functions in extracting temporal information from time stamped document sets.

6 Conclusion

A time decomposition of a time stamped document set is often constructed to explicate the temporal information hidden in the document set. Measure functions assign a numeric value to keywords such that the value captures the significance of the keyword in a document set. Measure functions are crucial in identifying keywords significant in a document set and consequently in a time interval/decomposition. In this paper, we defined two measure functions based on the notion of entropy. The interval entropy measure function determines a keyword to be significant if its occurrence frequency is higher than other keywords in the document set. The keyword entropy determines that a keyword is significant in an interval if it has higher occurrence frequency in that interval when compared to other time intervals. The effectiveness of the measure functions is studied by applying them to a subset of Reuter’s news articles and a subset of Enron Email messages. Several optimal time decompositions of the document sets are constructed and quantitative metrics such as size of information content and information loss were measured. The measure functions were very effective in identifying keywords that occur with a relatively high frequency in a time interval or those that have non-uniform occurrence during the time period associated with the document set. The information loss of optimal

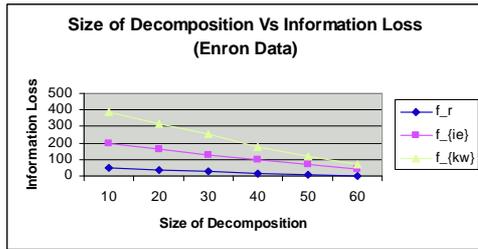


Figure 4: Size of Decomposition Vs Information Loss

time decompositions constructed with these measure functions falls as the size of the time decomposition increases.

References

- [1] J. Allan, V. Lavrenko, D. Malin, and R. Swan, "Detections, Bounds, and Timelines: UMass and TDT-3", *Proc. of the 3rd Topic Detection and Tracking Workshop*, 2000.
- [2] J. Allan, J. G. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic Detection and Tracking Pilot Study: Final Report", *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.
- [3] P. Chundi and D. J. Rosenkrantz, "Constructing Time Decompositions for Analyzing Time Stamped Documents", *Proc. of the 4th SIAM International Conference on Data Mining*, 2004, 57-68.
- [4] P. Chundi and D. J. Rosenkrantz, "On Lossy Time Decompositions of Time Stamped Documents", *Proc. of the ACM 13th Conference on Information and Knowledge Management*, 2004.
- [5] P. Chundi, R. Zhang, and D. J. Rosenkrantz, "Efficient Algorithms for Constructing Time Decompositions of Time Stamped Documents", *16th International Conference on Databases and Expert System Applications (DEXA)*, 2005.
- [6] P. Chundi and D. J. Rosenkrantz, "Information Preserving Decompositions of Time Stamped Documents", *Accepted to the Journal of Data Mining and Knowledge Discovery*.
- [7] T. M. Cover and J. A. Thomas, "Elements of Information Theory", John Wiley and Sons, NY 1991.
- [8] J. Kleinberg, "Bursty and Hierarchical Structure in Streams", *Proc. of the 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 2002.
- [9] J. Kleinberg, "Temporal Dynamics of On-Line Information Streams", In *Data Stream Management: Processing High-Speed Data Streams*, (Edited by M. Garofalakis, J. Gehrke, R. Rastogi), 2005.
- [10] B. Lent, R. Agrawal, and R. Srikant, "Discovering Trends in Text Databases", *Proc. of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD)*, 1997.
- [11] S. Roy, D. Gevry, W. M. Pottenger, "Methodologies for Trend Detection in Textual Data Mining", *Proc. of the Textmine 2002 Workshop*, SIAM Intl. Conf. on Data Mining, 2002.
- [12] R. Swan and J. Allan, "Automatic Generation of Overview Timelines", *Proc. of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2000, 49-56.
- [13] R. Swan and J. Allan, "Extracting Significant Time Varying Features from Text", *Proc. of the 8th International Conference on Information and Knowledge Management*, 1999, 38-45.
- [14] R. Swan and D. Jensen, "TimeMines: Constructing Timelines with Statistical Models of Word Usage", *Proc. KDD 2000 Workshop on Text Mining*, 2000.