Theme Encapsulation and Content Framework Implementation with Enhanced Lexical Chaining

A.V.Seetha Lakshmi
Asst.Prof, Department of IT
G.T.N Arts College, Dindigul, India

Dr.S.P.Victor
Head,Department of Computer Scienc
St.Xavier's College (Autonomous) ,PalayamKottai, India

Abstract—A developing event which contains many related events and activities is defined as a topic. There are many difficulties in exploration as there are sequences of documents published by different authors for the particular keyword. The convenience of storage in interne also adds to greater difficulty. The phenomenal growth of the internet has made the users to read the entire contents and conclude what exactly present in the document. In this paper, we present a comparative study of summarization in which the core content is summarized in the chronological order. The summarization process involves four steps 1. The original text is segmented, 2. Matrix calculation/ Lexical chain construction of the segmented data 3. Strongly related texts are identified, 4. Significant sentence are extracted. The extracted sentences are associated to find the temporal closeness with the help of evolution graph. The process of summarization based on the matrix calculation and lexical chains are compared based on the effectiveness of the text summarized and time taken to produce the accurate result.

Keywords— Data mining, Text mining, Encapsulation, Lexical chains, Matrix Calculation, Comparison, Summarization

I. INTRODUCTION

The internet provides an abundant source of information through the number of documents. The improvisations in the technology have paved the way for efficient search requests to satisfy the keyword search request. But, still readers have much difficulty in obtaining the required document from the abundant resources. The time related events add more controversy to the situation. The project, "Topic Detection and Tracking" initiated by the DARPA (Defence Advanced Research Project Agency) defines the topic as the *"semantic event or activity along with all directly related events and activities"*. In this paper, a comparative study of the two computational algorithm that detects all topics and track related documents from several documents using the keyword search technique is produced. A prominent text mining research pattern known as Topic anatomy is used to summarize the essential document in chronological order. It involves three major tasks such as *theme generation, event segmentation and summarization, evolution graph construction.*

The process of condensing a source text into shorter version by preserving the content information is known as summarization. The summarization process serves several goals such as survey analysis of a scientific field, quick notes on the general topic of the text, etc. There are two types of summaries namely informative summary and indicative summary. The quality informative summary can be produced only by the full understanding of the text. The quick indicative summaries are used to decide whether the text is worth reading and are easy to be produced. These indicative summaries are produced to select the particular topic for reading in chronological order.

In these techniques, the first task is to identify themes of the topic from the several document that are related. Each document may reflect its importance than the other while reading it. These events must be defined in a unique way. Later, the task of event segmentation and summarization process which extracts the events and presents the event in the chronological order is done. While our system TECF(Theme encapsulation and content framework) using lexical chains [2] and matrix calculation [1] detect the core content of the document in the effective manner. The source text needs to be integrated with the text representation to produce quality summaries. In this paper, we describe how matrix calculation method is effective than the lexical chains to identify highly reflected theme than the other themes..

II. RELATED WORK

A. Text Segmentation

The document related to the topic is divided into segments that are related to the topics. Non overlapping segments are formed in this technique. The segmentation can be classified into two types based on the input text as story boundary detection and document sub topic identification. The text stream is given as input to the story boundary detection. In general, cue phrases are used to identify the boundaries between the documents. For example, there are no distinct boundaries between the documents from online news documents. In document sub topic identification, with a single input document the blocks of document that are relevant to certain sub topic are identified. The cue phrase approach does not suit this technique as the subtopics in the document are similar. Therefore, the cue phrases for the document about the subtopic boundaries do not exist virtually. For example, Search engine can retrieve the documents and return the most relevant blocks segmented from the search results. The document can be decomposed into a set of consecutive sentences and the word usage in every block is analyzed in finding the subtopic boundaries. The major problem in this technique is that the block interrelationships cannot be determined with the information in the block.

Brants et al and Choi et al [3] to enrich the information in a consecutive set of sentences applied the latent semantics concepts. The training data is used to create a domain dependent construct in this method. Blei and Moreno [4] utilized Hidden Markov models to detect the subtopics of document and used it to model as states in an HMM. In this method, every document is treated as a series of blocks, which is used to calculate the best state transition. When two successive states in the best state transition sequence are different, then the boundary occurs in the documents.

Ji and Zha[5] proposed a domain-independent segmentation method that models the block of the document using a square matrix. It considers the matrix as a gray scale image. then some image processing method is applied to sharpen the boundaries in the image .Finally the significant and diagonal segments are selected as a block of the document.

B. Text Summarization

Text summarization creates one or more documents that capture the list of documents automatically. A document's content may consist of many themes, so generic summarization methods are used to extend the summary diversity to provide wider coverage of the content of the documents[6]. In text summarization, the informative sentence is extracted from the actual documents by composing the summaries. Extraction-based text summarization methods can be classified as supervised and unsupervised. In supervised methods, the document is summarized by labeling the sentences of the document as either informative or non informative. Shen et al[7] proposed a supervised summarization method that uses conditional random fields (CRF) to train a classification model. It calculates the information of the sentences. Top ranked sentences are selected as a summary with suitable training corpora. The supervised summarization methods perform as well as unsupervised summarization methods. Domain-dependency is a drawback of supervised summarization methods. The trained summarization model is specific to a certain document domain. Deploying a supervised summarization method in a new domain involves explaining another manual training corpus and it is a time consuming task. In general, the number of studies show that inter agreement between explanator is low, which affects the quality of the training corpus and the acquired summarization model .We recently proposed that the summarization method are unsupervised. Next, we consider the method and discuss the limitations of applying them to the topic anatomy task. Gong and Liu applied Singular value Decomposition (SVD) to a document. The term-sentence association matrix can be used to perform extraction-based generic summarization. This method uses the decomposed singular vectors for the themes of the document and composes diverse summaries by selecting the informative sentences from important themes. Nomoto and Matsumoto [8] proposed the X means algorithm ,which is used to find the sentences that contains more useful information from the clusters. Allan et al temporal summarization method [9] processes topic sentences in a chronological order. This method weights the information of the sentences depending on the usefulness and novelty of the sentences. A sentence is useful for readers to comprehend a topic if its content is similar to the main themes of the topic .To avoid the extraction of redundant summary sentences, the sentence should be different to all previously extracted sentences. Nowadays, graph based summarization methods are used to model the relationships between the sentence and terms in the document. This model considers a sentence informative if it connects with many informative terms and the reinforcement procedure updates the informative scores of the terms and sentences. Finally, the summaries are composed by selecting the informative sentences.

Erkan and Radev[10] represent the set of documents as a graph in the sentence are represented as nodes and the edges connects the content similarity between the sentences. A sentence is more informative, if it connects with many sentences, hence, the connected sentences are also informative. From the informative scores of the sentences, the informative sentence can be taken as the summary. Topic summarization differs from existing text summarization because of its temporal properties .The topic summaries should describe the storylines of the topics.

III. IMPLEMENTATION

A. *Theme Model*

The major tasks involved in the process are crawling and extraction, Matrix Calculation or Lexical chain computation, Event segmentation and summarization and finally the Story-boundary detection.

B. *Crawling and extraction*

The crawling is the basic operation performed in every web search Engines. The process of crawling takes place with the help of crawler. The crawler is a program that is already developed in the search engine. It isolates the citation from the World Wide Web (WWW) with respect to the particular theme through which we get the backlink. *[16].*

The processing steps involved in crawler are as follows,

i) The seed lists that are maintained by the crawler should be cleared first to make sure that the old data will not be provided to the user.
ii) Every webhosts has an IP Address, which is determined in this step.
iii) The crawler can download the document based on the user's search.
iv) The backlinks present in the document are extracted by the crawler.
v) The user can perform any operations on the downloaded document.
vi) While processing the document, the user may switch on to other citation and if the citation is new to the user, such citations are added to the seedlists .
vii) The process is repeated from (i) for every query requested by the user.

C. *Extraction*

We adopt two types of extraction called BLOCK EXTRACTION and THEME EXTRACTION.

Block extraction

In this, the blocks (the paragraphs) are extracted from the Original document using Indexing Technique**.** To ensure the chronological order, the blocks are extracted from the original document. While indexing, we assign index value to each tag used in source code. With that index value, we can identify the paragraph tag(i.e)<p> and with this, the paragraph are extracted.

Extracting blocks from the inbound links in the websites. In this block extraction, extracting blocks and image are extracted for the topic. We obtain n number of blocks from the web sites.

Theme extraction

The Themes (sentences) from the identified block are extracted using the Cue Phrase Identification Approach. The themes are extracted to reduce the complexity involved in Lexical chaining. After extracting blocks, Events are extracted from the blocks. A set of events are extracted for every blocks.

The cue phrase identification extracts the sentence by identifying the full stop at the end of each sentence.

D. *Matrix Calculation*

The matrix calculation for every themes which are extracted from the blocks of the document. This matrix should be orthogonal and n x n symmetric matrix.Calculating Eigen Values and Eigen Vector for every event from the blocks.

The eigenvalue equation for a matrix A is $Av - \lambda v = 0$ which is equivalent to $(A - \lambda I)v = 0$ where I is

the nxn identity matrix. It is a fundamental result of linear algebra that an equation Mv=0 has a non-zero solution v if and only if the determinant det(M) of the matrix M is zero. It follows that the eigenvalues of A are precisely the real numbers λ that satisfy the equation $\det(A - \lambda I) = 0$

By calculating the Eigen vector for events, we are getting a effective events for the particular topic from the peak values of the eigen vector[14]. Eigen vector with peak value is taken as a worthy event.The input to the matrix is taken as integer. This is done through stop word removal andsoundix.

- **Stopword Removal** In this, the frequently arriving words such as is, the, for, on etc are removed to make the matrix calculation less complex. This also helps to reduce the database spaces.

- **Soundix** In this, we arrive the integer value for the theme . With this value, we compare the relevancy between the theme and the topic. Eg.Speech recognition.

For any n * n symmetric matrix A of rank r, there exists a diagonal matrix D and an orthonormal basis V for R^n such that $A = VDV^{-1}$, where $V = v_1, v_2, \ldots, v_n$ consists of the eigenvectors of A; and the diagonal entries of D satisfy
$d1,1 \geq d2,2 \geq \ldots\ldots \geq dr,r > d_{r+1,r+1} = \ldots = d_{n,n} = 0$, which are eigenvalues corresponding to the respective columns of V.

Eg: <u>Topic summarization and content anatomy</u>

Topic summarization helps to identify the core parts of the document in the topic.

STOP WORD REMOVAL:

Topic summarization helps to identify the core parts of the document in the topic.

No. of content word in the theme=8

$\sqrt{8}=2.828$, rounding-off=3

Therefore, we arrive 3 x 3 matrix such that

Example:

$$A = \begin{bmatrix} 3 & 6 \\ 1 & 4 \end{bmatrix}$$

$$A - \lambda I = \begin{bmatrix} 3-\lambda & 6 \\ 1 & 4-\lambda \end{bmatrix}$$

$=(3-\lambda)(4-\lambda)-6$

$=12-4\lambda-3\lambda+\lambda^2-6$

$=\lambda^2-7\lambda+6$

$=(\lambda-6)(\lambda-1)$

$\lambda=6,1$ -> **Eigen Values**

E. Lexical chaining

The calculation of lexical chains can be done by two algorithms by Hirst et al and Starimind. These algorithms uses the WordNet lexical database which are represented through synonym sets (set of all words sharing common sense). For example two senses of "*computer*" are represented as: *{*calculator, reckoner, figurer, estimator, computer*}*. Similarly more than 118,000 word forms are available in the WordNet. The Words are linked through semantic relations like synonymy and hyponymy. Polysemous words appear in more than one synsets (synonym sets).

Three steps to construct the lexical chains:
1. Select a set of candidate words;
2. For each candidate word, find an appropriate chain relying on a relatedness criterion among members of the chains;
3. If it is found, insert the word in the chain and update it accordingly.

Relatedness of the words should be tenacious by the distance between their occurrences. Three kinds of cognations such as extra-vigorous (between a word and its repetition), vigorous (between two words connected by a Word-Net cognation) and medium-vigorous when the link between the synsets of the words is longer than one subsists. The candidate word is inserted in the congruous sense and the chain is updated if the chain is found else the incipient chain is engendered and inseted in the wordNet.For example:

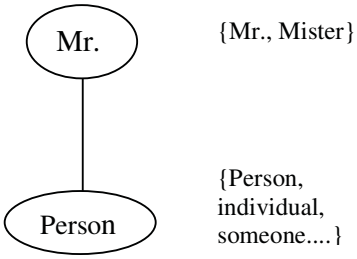


Fig:1: Step1

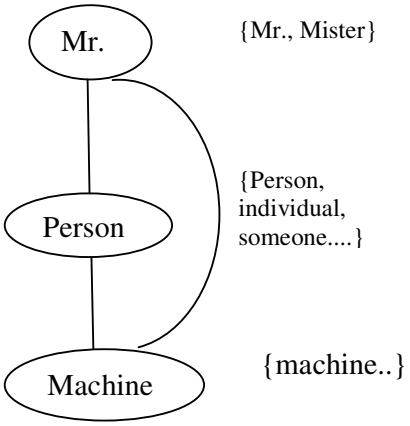


Fig:3: Step 3

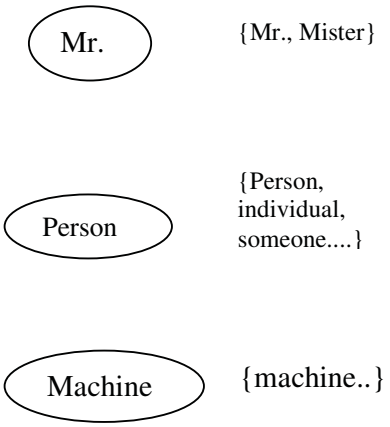


Fig:2: Step 2

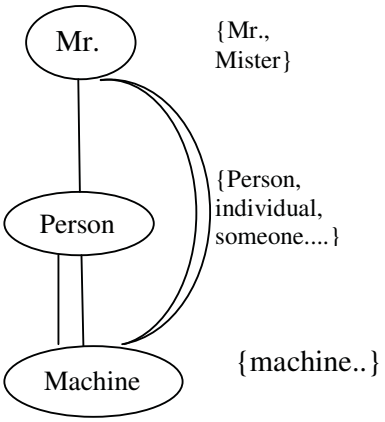


Fig:4: Step 4

Once the lexical chain is constructed, the summaries must be built using lexical chains. The chain length, distribution of text in the span, etc are identified to build the chain strongly. After the selection of strong chains, the next step is to extract full sentences from the original topic based on the chain distribution. Three alternative method can be used for this technique.

1) For each chain in the summary representation, choose the sentence that contains the first appearance of a chain member in the text.
2) For each chain in the summary representation, choose the sentence that contains the first appearance of a representative chain member in the text.
3) For each chain, find the text unit where the chain is highly concentrated. Extract the sentence with the first chain appearance in this central unit.

Concentration is computed as the number of chain members occurrences in a segment divided by the number of nouns in the segment.

F. Event Segmentation and summarization

A unique feature of summarization approach is the introduction of the event segmentation process to extract the semantic construct "event" before summarization.

G. Story Boundary Detection

Story boundary detection (or story segmentation) is used to identify where one story ends and the other story begins in a stream of text. It serves as a necessary precursor to various tasks, such as topic detection and tracking, information extraction, indexing, retrieval and summarization. A typical broadcast news retrieval system can locate the particular positions in a repository that match the user's query, but lack the ability of determining where the user-interested stories begin and end [20]. The data corpus consists of a collection of information, then using the classification techniques to create the boundary between two different news documents. The input from the event segmentation is detected core parts and from that, we have to identify the endpoints between the documents.

The story boundary detection is used to create the boundary between the two different news documents. Initially the end points are identified in the summarized documents and the boundary is created. This step is useful to display the news documents efficiently and helps the user to know different news.

H. Implementation of Enhanced Lexical Chaining

Enhanced Lexical chaining is the enhanced form of lexical chaining that improves the performance.
It is implemented as follows

Keywords:

Strip_tags	→ Strips white space characters from a sring
Search	→ Different webservers for finding the keywords to be searched.
Stopwords	→ Returns a string with all NULL bytes
explode	→ Split a string with the help of specific delimiter
path	→ Path to the wordnet database
enhanced_LexicalSearch	→ Function Name

```
function Enhanced_Lex_Search(search,searchTerm,stopwords)
{
blockSearch ← blockSearch('server',searchTerm)
    if(blockSearch='NULL')
    {
      return False
    }

// BLOCK EXTRACTION
foreach(blockSearch as block)
blockContent ← strip_tags(block,'content')


// THEME EXTRACTION
themeContents  ← explode('.',blockContent)
foreach(themeContents as theme)
        {
                findFlag=false
                foreach(searchTermArr as search)
                        {
                                If(strpos(theme),(search)!==false)
                                        {
                                          findFalg=true;
                                          singleTheme ← theme
                                        }
                        }
  }

// SEGMENTATION
if(findFlag)
        {
                foreach(searchTermArr as search)
                result ← execute('path'.search.'-synsn')
                if(result!= "NULL")
                {
                        Output['title'] ← block['title']
```

```
                    Output['url']   ← block['url']
                    Output['content']   ← theme
                    result[block['url']]  ← output
                }
}

// SUMARIZATION
if(result)
{
        resultSumar  ← array();
        foreach(result as key => resultSumar)
                {
                        returnresultSumar[key] = result[key]
                }
        return resultSumar
}

}
```

Algorithm for computing the result for Enhanced Lexical chain

1) foreach(searchTermArr as search)
2) result ← execute('path'.search.'-synsn')
3) if(result!= "NULL")
4) Output['title'] ← block['title']
5) Output['url'] ← block['url']
6) Output['content'] ← theme
7) result[block['url']] ← output

Definition:

1. The function Enhanced_Lex_Search performs a check to validate whether the search contains the content if so true is returned
2. If true, for each blockSerach the cntents are striped into blocks and returned for theme extraction.
3. In theme extraction, each theme within the keywords are extracted from the blockContent
4. For each searchTerm the theme is synchronised with the Wordnet to check for the closely related themes.
5. The theme or event is summarized as the output.

I. Comparison of Matrix Calculation method with the Lexical Chaining

Algorithm for computing Matrix Calculation:

1)serv.readthemes()
2)Initialize eigenvector
3) for (int i = 0; i <ds.Tables[0].Rows.Count; i++)
4)calc(dtb.ROW[i]theme,dtb.ROW[i]topic)
5)bind() values
6)page index changing()

Algorithm for computing Lexical chains:

Start.
For all candidate words do
Expand the words into possible senses (S_1, S_2,….S_n).
Determine the "offset " for each sense in WordNet .
End for
For all senses do

Insert the sense into the respective element of the synsetlD list
If inserted synset has relations with already inserted synsets then
Identify the relation and determine their score
End if
End for
For all relations do
Identify the chains compatible with the current relation
If compatible chain is found then
Insert into chain by looking out for repetition
Update the chainscore
Else create a new chain for the relation.
End if
End for
Sort the chains in descending order based on the chainscores
For all chains do
For all chainmembers do
If chain member already assigned a sense then
If assigned sense is not equal to the current chainmember sense then
FLG ← FALSE
End if
else
Assign the chain member the sense temporarily
End if
End for
If FLG equals to FALSE then
Discard the chain
else
assign the chain members their respective senses from the temporarily stored values
retain the chain
end if
end for
Stop

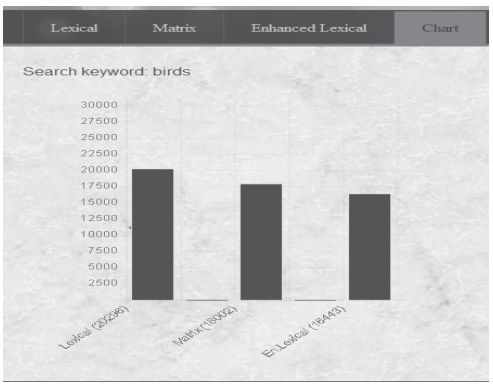
Comparative Study

The algorithm to compute the lexical chaining seemed to be more complex than the matrix calculation method and Enhanced Lexaical chaining. The word meaning varies with each thesaurus and so the effective way of constructing the chain was not achieved unless a constant method such as WordNet is followed for all documents. All the methods displayed accurate result at same probability.

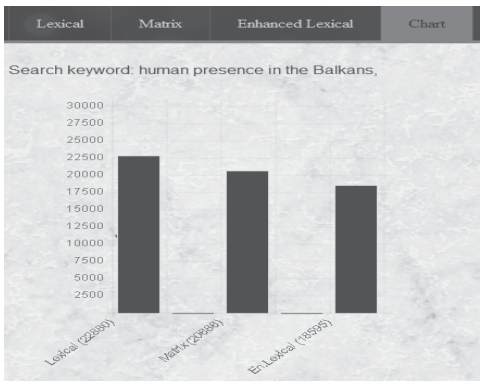
On the effective analysis of all the methods, lexical chaining was time consuming with respect to the matrix calculation method and Enhanced Lexical chaining. The analysis of this comparative study are plotted in tabular column and graph as follows

Keywords	**Lexical chains**	**Matrix Calculation method**	**Enhanced Lexical chaining**
Birds	20296ms	18002 ms	16443 ms
Human presence in the Balkans	22880 ms	20886 ms	18595 ms

Table:1



Graph: 1



Graph: 2

The time taken by all methods for each keyword search are shown in the above graph and tabular column. The Graph:1 denotes the time taken for the Keyword search birds. It denotes that the time taken by lexical chaining method is 20296ms whereas for matrix calculation its 18002ms and for Enhanced Lexical chaining its 16443ms. Similarly, Graph2 denotes the time taken for the keyword search human presence in the Balkans. Lexical chaining method took 22880ms and Matrix method took 20886ms and Enhanced lexical chaining took 18595ms. Thus the results of both the methods are compared effectively.

IV. CONCLUSIONS

Many news documents related to same topic are posted by different authors and their opinions vary during the topic life span. The summarization method is used to help the user to obtain the news from different documents.

In this paper, we have presented the implentation of Enhanced Lexical chaining and a comparative analysis on theme encapsulation and content frame work (TECF) using matrix calculation and lexical chains and Enhanced lexical chaining, which extracts the themes, events and connects the associated events to form evolution graph. Matrix calculation method is better choice than the Lexical chaining method as matrix calculation was less time consuming than the Lexical chaining method and Enhanced Lexical chaining is better than Matrix method as it is faster.

Thus, Enhanced Lexical chaining method is the best method suited for Theme encapsulation and content framework.

REFERENCES

[1] A.V.Seetha Lakshmi and Dr. S.P.Victor," TECF:Accomplishment of Content Framework Tactic to Temporal Theme Condensation", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 11, November 2013.

[2] A.V.Seetha Lakshmi and Dr. S.P.Victor, " Theme Encapsulation and Content Framework using Lexical Chaining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 3, March 2014.

[3] Chien Chin Chen and Meng Chang Chen,"*TSCAN: A Content* Anatomy Approach to Temporal Topic Summarization",IEEE Transactions onKnowledge and Data EngineeringVolume:24Issue:1

[4] D.M. Blei and P.J. Moreno, "Topic Segmentation with an Aspect Hidden Markov Model," Proc. 24th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval,pp. 343-348, 2001.
[5] X. Ji and H. Zha, "Domain-Independent Text Segmentation Using Anisotropic Diffusion and Dynamic Programming," Proc.26th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 322-329, 2003.
[6] Y. Gong and X. Liu, "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis," Proc. 24th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 19-25, 2001.
[7] D. Shen, J.T. Sun, H. Li, Q. Yang, and Z. Chen, "Document Summarization Using Conditional Random Fields," Proc. 20th Int'lJoint Conf. Artificial Intelligence (IJCAI '07), pp. 2862-2867, 2007.
[8] T. Nomoto and Y. Matsumoto, "A New Approach to Unsupervised Text Summarization," Proc. 24th Ann. Int'l ACM SIGIR Conf.Research and Development in Information Retrieval, pp. 26-34, 2001.
[9] J. Allan, R. Gupta, and V. Khandelwal, "Temporal Summaries of News Topic," Proc. 24th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 10-18, 2001.
[10] G.Erkan and D.R. Radev, "LexRank:Graph-Based Centrality as Salience in Text Summarization," J. Artificial Intelligence Research, vol. 22, pp. 457-479, 2004.
[11] J. Kleinberg, "Bursty and Hierarchical Structure in Streams," Proc.Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and DataMining, pp. 91-101, 2002.
[12] R. Nallapati, A. Feng, F. Peng, and J. Allan, "Event Threading within News Topics," Proc. 13th ACM Int'l Conf. Information and Knowledge Management, pp. 446-453, 2004.
[13] C.C. Yang and X. Shi, "Discovering Event Evolution Graphs from Newswires," Proc. 15th Int'l Conf. World Wide Web, pp. 945-946,2006.
[14] A. Feng and J. Allan, "Finding and Linking Incidents in News,"Proc. 16th ACM Conf. Information and Knowledge Management,pp. 821-830, 2007.
[15] R. Swan and J. Allan, "Automatic Generation of OverviewTimelines," Proc. 23rd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 49-56, 2000.
[16] *Web Crawlers & Hyperlink Analysis.*Albert Sutojo. CS267 – Fall 2005. Instructor :Dr. T.YLin.
[17] A. Nenkova, L. Vanderwende, and K. Mckeown, "A Compositional Context Sensitive Multi-Document Summarizer: Exploringthe Factors that Influence Summarization," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval,pp. 573-580, 2006.
[18] L E. Spence, A.J. Insel, and S.H. Friedberg, Elementary Linear Algebra, a Matrix Approach.Prentice Hall, 2000.
[19] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval.Addison Wesley, 1999.
[20] Broadcast News Story Boundary Detection Using Visual, Audio And TextfeaturesMaryamDaneshi, Matt Yu
[21] Kleinberg,J. "Authoritative Sourcesin a Hyperlinked Envi- ronment," in Proceedings of the ninth annual ACM-SIAMsymposiumon Discretealgorithms,1998,668-677.