Winter 12-19-2001

Coherence Identification of Business Documents: Towards an Automated Message Processing System

Samuel W.K. Chan

COHERENCE IDENTIFICATION OF BUSINESS DOCUMENTS: TOWARDS AN AUTOMATED MESSAGE PROCESSING SYSTEM

Samuel W.K. Chan
Department of Decision Sciences and Managerial Economics
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong

ABSTRACT

This paper describes our recent efforts in developing a text segmentation technique in our business document management system. The document analysis is based upon a knowledge-based analysis of the documents' contents, by automating the coherence identification process, without a full semantic understanding. In the technique, document boundaries can be identified by observing the shifts of segments from one cluster to another. Our experimental results show that the combination of the heterogeneous knowledge is capable to address the topic shifts. Given the increasing recognition of document structure in the fields of information retrieval as well as knowledge management, this approach provides a quantitative model and automatic classification of documents in a business document management system. This will beneficial to the distribution of documents or automatic launching of business processes in a workflow management system.

INTRODUCTION

Last decade has witnessed one of the dramatic progresses in the area of message understanding processing. Buoyed by the great demand in information retrieval, computational linguists have found themselves at the centre of an information revolution ushered in by this Internet age [1, 3, 15]. In respect to the well-established information superhighways and the challenges of the content-based retrieval, message understanding processing is certainly a key technology for building up information systems in the next generation. Unlike the current relatively crude search engines that retrieve long lists of documents of often questionable relevance, the needs which systems can "understand" both the user queries and the semantic information possessed in textual databases are still mounting.

Message understanding is a process aimed at converting a linear sequence of a text into coherent wholes. Most linguists agree that a primary activity during message understanding is connecting the phrase and sentence that is currently being read with the contents of the immediately preceding sentences. If a reader is able to make such a connection, then coherence is said to achieve. While it is important for readers to be able to trace continuities in the entities under discussion, it is equally important to locate and understand the breaks or turns in a text. Given the increasing recognition of text structure in the fields of information retrieval in un-partitioned text, an implementable quantitative model in coherence identification becomes inevitable.

The recent literature on textual information systems has begun to respond to this challenge. Morris & Hirst suggest that the discourse structure of a text can be determined through an analysis of lexical cohesion [9]. Using hand coding, they use a thesaurus to identify chains of related words across sentences. Breaks in these lexical chains trend to indicate structural elements in the text, such as changes in topics and the writer's intentional structures. Followed the work of Youmans [24], the concept of text window is used within which they compute a lexical cohesion function. By moving the window over the text, they form a linear plot of the lexical cohesion as a function of the word position. A discourse boundary is assigned if the value falls below a threshold. In a similar vein, Kozima proposes a lexical cohesion profile as a quantitative indicator of marking text boundaries [6]. The profile is a record of lexical cohesiveness of words in a window that moves forward word by word on a text. Since a coherent text tends to be lexically cohesive, the local cohesiveness suggests coherence in the text. More recently, Yaari segments text into a hierarchical structure, identifying subsegments of larger segments [23]. Ponte & Croft use lexical co-occurrences to expand the number of terms for matching [14]. Litman & Passonneau present an algorithm that use decision trees to combine multiple linguistic features extracted from corpora of spoken text, including prosodic and lexical cues [7]. However, their approach is lacking in features related to lexical cohesion. Nomoto & Nitta detect coherence through patterns of text co-occurrence [10]. They adopt the *saliency* factor as one of their weighting policies. Hearst segments expository texts into multiple paragraphs of coherent discourse units [4]. A cosine measure is used to gauge the similarity between constant-size blocks of morphologically analyzed lexical items. However, all the above approaches rely on a single model and take the basic assumption that linguistic knowledge is homogeneous. This assumption can be very misleading while considerable heterogeneity and diversity can be found in linguistic knowledge. They do not flexible enough to work in message understanding environment which is characterized by almost infinite variability.

Identification of coherence is a multifaceted process involving most linguistic knowledge, at least lexical repetition, as well as semantic overlapping. These knowledge sources cooperate in a more or less synchronized way. However, the information provided by each knowledge seems independent

of and has not much bearing on each other. None of these can fully identify the coherence or provide a general solution to the identification, but each will provide clues to solve the problem. Central to our approach is taking lexis as a starting point for coherence identification. The main objectives of this research are (i) to investigate patterns of coherence in expository texts in order to test hypothesis about the textual continuity; (ii) to devise a measure in order to analyze the interrelations between each segment; (iii) to formulate a computational model and an objective measure in analyzing coherence; (iv) to propose and implement a method for the segmentation of texts into thematically coherent units.

The paper is organized as follows. In the following section, we shall first explain our basic formalism in document modeling. A brief discussion of the orthogonal linguistic knowledge, which includes lexical preferences and token saliency factor, will appear in the next section. We shall also explain how the system can identify the coherence from the piecewise orthogonal knowledge sources. In order to demonstrate the capability of our system, a simulation is delineated followed by a conclusion.

COHERENCE ANALYSIS

A document is composed of a number of paragraphs, each of which is made up of a number of segments. A discourse segment may be a group of sentences or long phrases. Given that our intention is to explore the means by which various linguistic factors link sentences, it is necessary to have a formalism for representing the links that will accurately reflect the non-linear complexity of a document and, at the same time, permit us to handle and interpret them conveniently. In our consideration of how a document structure is expressed, we have already established a discourse network that is employed to represent the inter-sentential relationships existing among the segments.

[DEFINITION 1]
A discourse network ***D*** is defined by a set of discourse segments, which stands in functional relations to each sentence in the discourse. The discourse network is represented as a graph characterized by a 5-tuple [2, 13, 5].

$$\boldsymbol{D} = \langle G, T, A, E, W \rangle \quad \text{where}$$

- G is a finite set of the discourse segments composing the document.
- T is a finite set of lexical items (hereafter, called *token*) composing the discourse segments.
- A is a set of arcs representing the inter-sentential relations amongst the discourse segments.
- E is a set of weights of the arcs.
- W is a function $W: A \rightarrow E$ which assigns lateral weights to arcs.

In our discourse network, the lateral weights between the arcs among the discourse segments are defined by heterogeneous linguistic knowledge. Let g_i, $g_j \in G$ be two discourse segments in the discourse network ***D***, each representing a different segment. If both of these segments are interrelated, the connection between them, i.e., W_{ij}, is assigned a large positive weight. On the other hand, it is reasonable to assume that syntactic function words do not denote new topics, whereas new semantic content words (nouns, main verbs, adjectives, and some adverbs) do. Given this assumption in our identification, a segment could be generated for a document simply by removing all function words from those tokens in the preprocessing.

One aspect of world knowledge essential to constructing the network is to know when two tokens in the segments are related. Several major types of relationships provide a document with coherence. First, lexical preference or similarity is crucial in solving many message understanding processing tasks [11, 17, 20]. It provides the inter-cohesive structure by relating its tokens to one another. In addition to the above lexical preference, we also adopt the token saliency factor, as corpus-based knowledge, which takes into consideration the frequency of occurrence of the processing token in the database [18]. We distinguish the semantics of an item from others through their co-occurrence across different documents in the document database. It can be regarded as an associate meaning relationship between regularly co-occurring tokens. In the following sections, we will describe how this heterogeneous linguistic knowledge can be utilized in building up the discourse network ***D*** and how our coherence identification can take advantage of them.

HETEROGENEOUS LINGUISTIC KNOWLEDGE

Lexical Preferences

While the syntactic constraints, though necessary, have little extension in semantic dimensions of the domain of analysis, it is well known that evaluating semantic similarity is crucial in solving many tasks in message understanding. An essential component of the lexical entry of a word is a definition of its meaning. Whittemore and his colleagues find lexical preferences to be the key to resolve ambiguity, however, arriving at an adequate representation of the meaning of a token is a notoriously difficult task [22]. They echo Taraban and McClelland [21, 8] who have shown that the structural models of message analysis are not in fact good predictors of human behavior in semantic interpretation. In this paper, in order to measure the lexical preferences among a message, we employ an *is-a* semantic net and argue that the shortest paths between any lexical items significantly correspond the semantic distance and semantic relatedness [16]. In our formalism, let X and Y be two lexical items represented by the nodes x and y respectively in an *is-a* semantic net, a measure of the lexical semantic relatedness between X and Y is given by

Distance(X, Y) = minimum number of edges separating x and y.

At the same time, the semantic similarity measure S between the lexical items, or tokens, is defined by:

$$S(x,y) := \begin{cases} \langle x,y \rangle & \text{if Distance}(X,Y) \le d_{max} \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

where $\langle x, y \rangle = [1+\text{Distance}(X, Y)]^{-1}$. Moreover, in order to compute the conceptual distances between every segment-pair, all pairwise combination between tokens in one segment g_i and tokens in every other segment g_j are generated. For

each pairwise combination, the following definition is used to create a metric over the segments in the discourse network.

[DEFINITION 2]
Let $g_i = \{w_{x1}, w_{x2}, ..., w_{xm}\}$ and $g_j = \{w_{y1}, w_{y2}, ..., w_{yn}\}$ be the two segments, the similarity components due to the lexical preferences are defined by:

$$s_S(g_i, g_j) = \frac{1}{\max(m,n)} \sum_{w_{xi} \in g_i} \max_{w_{yi} \in g_j} \{S(w_{xi}, w_{yi})\} \quad (2)$$

Token Saliency Factor

Token saliency factor is a weight function which computes the number of each token occurs in a document *tf* times the inverse logarithm factor of the number of documents that the token occurs in a large collection *idf* [18,19]. Based on the knowledge from corpus, one advantage of using token saliency factor is that the boundary of topics in a document can be distinguished by the coherence values of each segment pairs [10]. The saliency factor for each token in a document is defined as

$$r(w_i) = tf(w_i) \times idf(w_i) \quad (3)$$

where token frequency *tf* is the number of occurrences of a token w_i in a document D_i. Document frequency *df* is the number of documents in a collection of N in which the token w_i occurs. The saliency factor $r(w_i)$ is the product of *tf* and $\log \frac{N}{dfi}$, the inverse of *df* factor. When N is large and df_i is small, the token w_i is considered to be more important than others. However, if N is large and the df_i is large too, the token w_i is considered to be less important in the document. The frequent tokens that are concentrated in particular documents are considered to be more important than the other frequent tokens that occur evenly over the entire document collection. In other words, the saliency factor favors rare words than common words. Tokens that commonly found throughout a collection are not necessarily good indicators of saliency. As a result, their importance is downweighted. In our approach, in order to determine the coherence value between segments, the following definition is employed.

[DEFINITION 3]
Let $g_i = \{w_{x1}, w_{x2}, ..., w_{xm}\}$ and $g_j = \{w_{y1}, w_{y2}, ..., w_{yn}\}$ be the two segments, coherence value for the similarity between segments is calculated by a normalized inner product of the two text segments g_i and g_j, the similarity component due to the saliency factor is defined by:

$$s_{coh}(g_i, g_j) = \frac{2\sum_{i,j} r(w_i) \times r(w_j)}{\sum_i r(w_i)^2 + \sum_j r(w_j)^2} \quad (4)$$

It yields a value between 0 and 1 representing the coherence value between the segments.

COHERENCE IDENTIFICATION

The weight generated from all these two major principles are combined to form an overall lateral matrix W which represents the connection across each segment-pair.

$$W(g_i, g_j) = \alpha s_S(g_i, g_j) + \beta s_{coh}(g_i, g_j) \quad (5)$$

where α and β are the proportional constants.

A great concentration of weights is near the diagonal of the lateral weight matrix W which indicates there is high coherence at the neighborhood of each segment. Obviously, link concentration is a potential indicator of coherence. Most of the text segmentation techniques are based on the premise that the coherence should be lower where the topic changes. Our coherence identification turns to identify clusters generated from the lateral weight matrix. Segments with high coherence will form a cluster. Boundaries are detected through the shifts of discourse segments from one cluster to another. We make use of an orthogonal decomposition known as the Singular Value Decomposition (SVD), which is a generalization of the well-known eigenvalue decomposition. It has been used in the solution of unconstrained linear least squares problems, matrix rank estimation and canonical correlation analysis. In the remaining section, first, we shall explain what the SVD is and then demonstrate how it can be applied to identify the discourse boundaries.

[THEOREM 1]
Given a matrix $A \in \mathbf{R}^{m \times n}$, without loss of generality $m \geq n$ and rank $(A) = r$, then there exists orthogonal matrices $U \in \mathbf{R}^{m \times n}$ and $V \in \mathbf{R}^{m \times n}$ such that

$$A = U \Sigma V^T$$

$$\text{where} \quad \Sigma = \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix}$$

and $\Sigma = \text{diag}(\sigma_1, \sigma_2, .., \sigma_n)$, $\sigma_i > 0$ for $1 \leq i \leq r$, $\sigma_j = 0$ for $j \geq r + 1$ and

$$U^T U = V^T V = I$$

The first r columns of the orthogonal matrices U and V define the orthogonal eigenvectors associated with the r nonzero eigenvalues of AA^T and $A^T A$ respectively. The singular values of A are defined as the diagonal elements of Σ which are the nonnegative square roots of the n eigenvalues of AA^T. These matrices reflect a breakdown of the original relationships into linearly independent vectors or factor values.

In our application, the first step is to represent the interrelationships among the segments in the text, as defined in Eqn. (5), by an overall $m \times m$ lateral weight matrix W in which each row and column stands for a unique segment. Each entry, say W_{ij}, represents the weight in which the segment i is related to segment j and the entry subsumes the contribution coming from the lexical preferences and the token saliency. The SVD of the matrix W is then defined as the product of three matrices,

$$W = B \Sigma B^T$$

where the columns of B contains the eigenvectors of W and Σ is a diagonal matrix containing the eigenvalues in descending order:

$$\sigma_1 \geq \sigma_2 \geq .... \geq \sigma_n$$

The eigenvectors are normalized to have length 1 and orthogonal, which means that they satisfy the following condition: $B^T B = I$. Decomposing a regular matrix into a product of

three other matrices is not too interesting. However, if the first k ($<< m$) columns of the B matrix and the first (largest) k singular values of W are used to construct a rank-k of W via W_k, such that

$$W_k = B_k \Sigma_k B_k^T \qquad (6)$$

then the W_k constructed from the k-largest singular triplets of W is the closest rank-k approximation in the least squares sense. Using the Singular Value Decomposition with k is set to 2, our lateral weight matrix W is truncated into a new segment-by-segment matrix by multiplying the first two singular values of diagonal matrix Σ with the first two columns of singular vectors of the orthogonal matrix B.

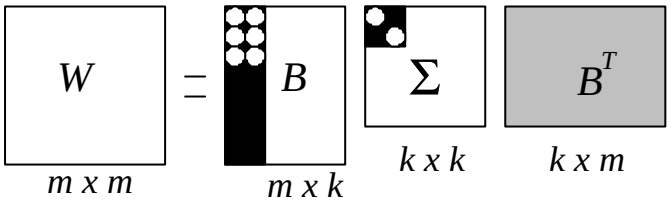

Figure 1: Mathematical representation of the matrix W

The advantage of using SVD is that the truncated matrix can approximate the higher-order structure in W representing the association of segments within the document. W_k is the best possible rank-k approximation of W in several senses, including the square root of the sum of squares of the elements. Another way to express this is that if we project onto the first k principal components, we have the most accurate rank-k reconstruction of the original data points in W_k. The truncated SVD matrix is used to show the high coherence relationships among the segments in the text. It also captures the most important underlying text structure in terms of the interrelationships among segments and removes the noise or variability that plagues the prominent coherence ties.

The major coherence patterns among the text segments can be represented geometrically under the decomposition. The result of our coherence identification using SVD is a k-dimensional vector space having a vector for each segment. These vectors have a geometric interpretation as they define points in a multidimensional space. In order to visualize the topic changes in a text, we represent each segment within the text in a two-dimensional Cartesian plane. We make use the first column of B multiplied by the first singular value λ_1 for the x-coordinates and the second column of B multiplied by the second singular values λ_2 for the y-coordinates. The proximity of segment vectors in the plane reflects the similarity in their coherence. As a result, clusters of segment vectors can be found in the two-dimensional Cartesian plane. The cluster means a group of segments that is linked by the two coherence factors as described in the last section. Segments within a cluster have a high coherence index and fully comprehensible. When more coherence ties can be found between the segment pairs, the pairs will be more contiguous in the two-dimensional plane and it is most unlikely that the segment boundaries lie among them. In other words, segments with high coherence or under the same topic will form a cluster while any topic shift in text can be detected by the shift of segments from one cluster to another.

IMPLEMENTATION AND RESULTS

In order to demonstrate the flavor of the analysis and what the technique has accomplished, a simulation is used to show the inter-relationships among segments within documents and to illustrate how the principles influence textual segmentation. Fifteen documents are selected in our implementation with a total of 476 segments and more than 14,000 tokens. The documents are extracted from several categories, including commerce, information technology, entertainment, health, education and sport. The total number of paragraphs obtained is 69. In order to ensure there are topic shifts between the paragraphs within a document, we mingle these 69 paragraphs together in order to produce 15 new documents by randomly selected 4 to 5 paragraphs for each new document. They are also under a constraint that no consecutive paragraphs are coming from the same original documents. In addition, with the assumption that function words do not denote much important meaning while semantic content words do, our document preprocessing first removes function words from the documents. At the same time, other relevant information, such as segment ID, segment-token number, and token ID are stored into a database. In order to represent the sole effect of each principle as described in last section, we demonstrate their outcomes one-by-one in a document with 23 segments. The newly generated document is composed of four major paragraphs which come from the journal *Harvard Asia Pacific Review, Asia Computer Weekly* and editorials of a local paper. All these paragraphs are under the same topic – *Hong Kong*, but certainly with different themes. This ensures that both coherence and paragraph boundaries can be clearly identified without any subjective judgement. Under the principle of lexical preference, every pair of segments is compared to find the number of same or similar tokens as defined in Eqns. (1) and (2). As more repetition among tokens can be found between the segments, this segment pair will have a higher coherence index.

Figure 2 shows the segment clusters formed under lexical preference after the singular value decomposition with k equal to 2. The number in the figure represents the corresponding segment in the document. The ovals indicate those segments that are likely close together and may be considered as coherent wholes under the same topic. The distance between segments in this two dimensional plane reflects the similarity among them. That is, the denser the segments appear in the ovals, the higher the similarity of the segments is.

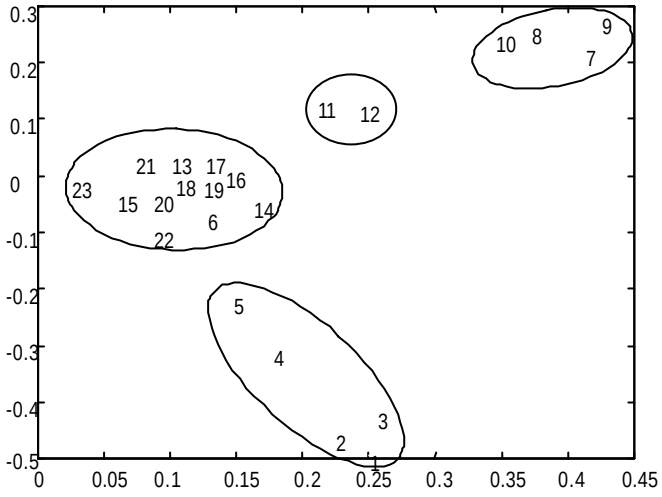

Figure 2: Segment cluster formed by lexical preference after the singular value decomposition with $k = 2$. Boundaries occur at discourse Segments 5, 10, 12 since the segments shift from one cluster to another.

Similarly, in token saliency factor, coherence is measured at every pair of segments as defined in Eqns. (3) and (4). Or more precisely, given a segment g_i and a block size n, what we do is to compare a block spanning from g_{i-n+1} to g_{i+n-1}. It is based on the idea that the number of token saliency links shared by segment pairs tends to increase as the distance between segments decreases. In representational terms, this means that there is a greater concentration of links near the diagonal of the lateral weight matrix.

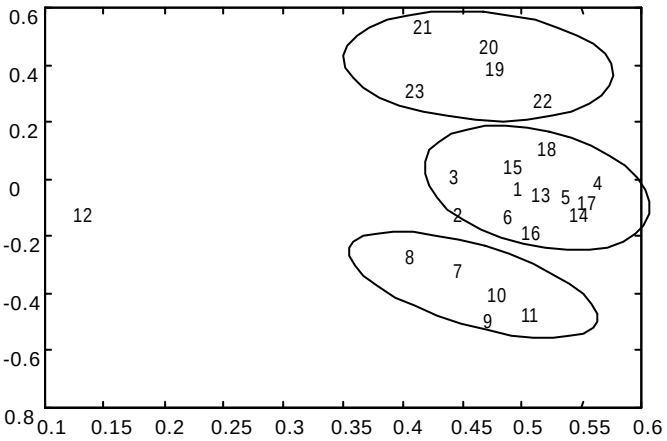

Figure 3: Segment cluster linked by token saliency factor after the singular value decomposition

Figure 3 shows the segment cluster produced by token saliency factor after the singular value decomposition. It can be observed that Segments 7-11, 19-23 are under different topics. However, the boundaries between the two clusters, 1-6 and 13-18 are totally unclear. Chaos appears in Segments 13-18 which seem to overlap with Segments 1-6. The application of the lexical preference and token saliency factor is supported by empirical studies of text structure, and each is consistent with general assumptions about the nature of document. However, as shown in Figures 2 and 3, it is clear that none of these principles are by themselves sufficient to have a reasonable solution to text segmentation. In the process of identification of coherence among the text segments, the reader must make a number of bridging inferences that do not solely rely upon either one. Given this situation, our working hypothesis described here is that all these heterogeneous knowledge sources must be applied simultaneously. Figure 4 shows the combined effect of the two principles such as lexical preference, as well as token saliency. The figure shows that Segments 1-5, 7-11, 13-18, 19-23 are clearly under different topics and the corresponding topic shifts occur at Segments 6-7, 12-13, 18-19. By investigating these results in turn, it is clear that the combined effect achieves the best result, although the lexical preference, among the two principles, shows an acceptable performance. One may expect that the performance will be deteriorated by the inappropriate links as more inter-relationships are added. However, the dimension reduction using singular value decomposition has demonstrated its capability by distilling the main gist or segment clusters in the noisy environment. This coarse segmentation provides the outlines and the gist of the text, omitting details and inconsistencies. This textual segmentation has obvious applications at the beginning of any summarization processes [12].

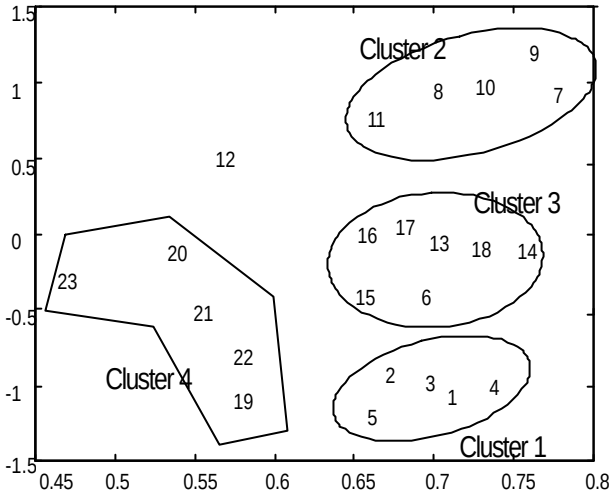

Figure 4: Segment clusters linked by the combined effect after the singular value decomposition

In the following excerpts, we begin by looking at how the clusters that identified are mapped onto the existing boundaries across the document. We also provide the lexical evidences, in the first two clusters, to illustrate how segments can be connected in a coherent series according to the semantic or pragmatic relations identified so far. The first cluster, from Segments 1 to 6 as shown below, emphasizes the technology challenge in Hong Kong.

(*Cluster 1, Segment 1*)
An economic greenhouse or protected technology zone is required to encourage both home grown technology initiatives in Hong Kong and China's fledgling technology industries. Hong Kong needs the political impetus to develop an Asian Silicon Valley, which would provide an appropriate environment for the research

(*Cluster 1, Segment 2*)

and development of products and services; this is turn would produce the trade and revenues for the future. Plans for an applied science and technology research institute have been announced by the Hong Kong government. This institute will provide the means and

(*Cluster 1, Segment 3*)
the impetus to link basic scientific research to its commercial applications and provide start-up support, such as low-cost office rent and access to international research facilities. Another Hong Kong government initiative is the establishment of a US$5billion innovation and technology fund to finance projects

(*Cluster 1, Segment 4*)
which will contribute to technical innovation in Hong Kong. A US$5 billion quality education fund has also been created to raise school standards through innovation and to facilitate technological collaboration between academic institutions and the business community.

(*Cluster 1, Segment 5*)
These funding initiatives create the foundation for economic success. Strong financing provides the leverage for research and development and secures the stability for venture partnerships, which will in turn attract foreign investment and expertise into local markets.

(*Cluster 1, Segment 6*)
Hong Kong is now poised to become the powerhouse of e-commerce. The proposed Hongkong.com Studio would leverage Hong Kong's core strengths in commerce and communication, stimulate job creation and wealth creation opportunities in both Hong Kong and China.

(Extracted from *Harvard Asia Pacific Review*, 3, 1, 1999, pp.40-41)

In addition to the appearance of the exact lexical repetition, such as the lexical items, *technology*, *research*, *impetus*, *initiatives*, *innovation*, *Hong Kong* in the first five segments, the lexical preference bonding can also be observed among the segments as shown in Table 1.

Lexical Preferences	Segment 1	Segment 2	Segment 3
Technology	*Silicon Valley*	*Applied Science*	*Technology/ Scientific*
Government	*Political*	*Government*	*Government*
Economy	*Economic*	*Trade/ Revenues*	*Commercial*
Fund	*N/A*	*Revenues*	*US$5 billion*

Lexical Preferences	Segment 4	Segment 5	Segment 6
Technology	*Technological*	*N/A*	*N/A*
Government	*Institutions*	*N/A*	*N/A*
Economy	*Business*	*Economic/ Investment/ Financing*	*Commerce*
Fund	*US$5 billion*	*Fund*	*Wealth*

Table 1: Links from lexical preferences among the segments in the first cluster.

While the exact lexical repetition may directly influence the token saliency factor as shown in Eqn. (4), the lexical preferences, as captured via our *is-a* semantic network shown in Eqn. (2), indicate the coherence as well as diversified sources of interrelations among the cluster. In fact, in the computational linguistic research, *collocation cohesion* describes the kind of ties created by lexical items that are related to each other only insofar as they tend to appear together in similar contexts. Our lexical preferences and saliency factor create various degrees of cohesiveness on the basis of the frequency of their occurrences and proximity in a text. This collocation cohesion also explains the early formation of Cluster 1 as shown in Figures 2 and 4. Similarly, Cluster 2 begins at Segment 7 and stretches up to Segment 11. It focuses on the notion of the Hong Kong movies in Hollywood. This document expresses the personal view of the author, a Harvard Professor of Modern Chinese Literature, on Asian influences in American popular culture. This explains why the personal pronouns, such as *I*, *my*, *we*, *our*, *us* can be found across the document. In addition, lexical items with part-of relation, such as *American*/*Washington*, *United States*/*Washington*, and various forms of contrast, such as *Global*/*Provincial*, also provide an unvarying, systematic semantic relationship with each other. Expressions of similarities or contrasts seem to be of a similar kind of intellectual operation. They both link up the segments within the cluster.

(*Cluster 2, Segment 7*)
I choose Hong Kong movies, my favorite subject, for a good reason. It's about time that Americans become exposed to foreign cultures, whether they come from elite or popular sources, because despite its global ambitions,

(*Cluster 2, Segment 8*)
the United States has become one of the most "provincial" countries in the world. Academically speaking, there's nothing new about Asian influence in America. My colleagues and I have been talking about it for several decades.

(*Cluster 2, Segment 9*)
However, our effort in studying and teaching Asian cultures has proven successful only on college campuses. Elsewhere, American mass media continues to overwhelm the general public with sensational images, simplistic stereotypes, and

(*Cluster 2, Segment 10*)
preconceived notions of what makes Asians tick. The business leaders fare a little better than the politicians in Washington, but their understanding of Asian cultures remains limited to politics and money, and

(*Cluster 2, Segment 11*)
does not focus on culture in any form. Thus the popularity of Chinese movies from Hong Kong is a phenomenon worthy of attention, for it gives us an example of how a specimen of that culture is making inroads into the American mainstreams.

(Extracted from *Harvard Asia Pacific Review*, 3, 1, 1999, p.30)

The lexical evidences which provide the possible links among the segment can be found as follows:
Hong Kong → Asian → Chinese
I → My → Our → Us
Culture → Culture → Culture
Americans → United States → Washington → American
Movies → Mass Media → Movies
Global → Provincial

CONCLUSION

In this research, the modeling we put forward is to employ a novel approach which establishes a network of interrelations among segments in a document. Coherence between linguistic items is reflected by using various linguistic clues modeled in our discourse network. The process of text segmentation, from a microscopic point of view, can be regarded as a process of assigning weights between the text segments. In order to exaggerate the coherence effect, we have presented a method for segmenting texts into thematically coherent units using the SVD technique in matrix computation. Our initial discourse network is subjected to a singular value decomposition which is interpreted as a particular transformation of a given set of weights into a set of segment clusters. This novel approach, different from any others, not only provides more sophisticated text segmentation by reducing the noise but also provides a clear visual effect in the analysis.

ACKNOWLEDGEMENTS

The work described in this paper was partially supported by two grants from the Research Grants Council of the Hong Kong SAR, China (Project Numbers: CUHK 1196/99E and CUHK 1221/00E).

REFERENCES

[1] C.C. Byrne and S.A. McCracken, An adaptive thesaurus employing semantic distance, relational inheritance and nominal compound interpretation for linguistic support of information retrieval, *Journal of Information Science* 25(2) (1999) 113-131.

[2] S.W.K. Chan and J. Franklin, A Brain-State-in-a-Box Network for Narrative Comprehension and Recall, *Proceedings of IEEE International Conference on Neural Networks*, Washington, D.C., (1996) 694-699.

[3] J. Hartley, M. Sydes, and A. Blurton, Obtaining information accurately and quickly: Are structured abstracts more efficient? *Journal of Information Science* 22 (5) (1996) 349-356.

[4] M.A. Hearst, Texttiling: Segmenting text into multi-paragraph subtopic passages, *Computational Linguistics* 23 (1997) 33-64.

[5] W. Kintsch and T.A. van Dijk, Toward a model of text comprehension and production, *Psychological Review* 85(5) (1978) 363-394.

[6] H. Kozima, Text segmentation based on similarity between words, *Proceeding of the thirty-first Annual Meeting of the Association for Computational Linguistics* (1993) 286-288.

[7] D.J. Litman and R.J. Passonneau, Combining multiple knowledge sources for discourse segmentation, *Proceedings of the thirty-third Annual Meeting of the Association* (1995).

[8] K.R. McKeown, *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text. Studies in Natural Language Processing* (Cambridge University Press, 1985).

[9] J. Morris and G. Hirst, Lexical cohesion computed by thesaural relations as an indicator of the structure of text, *Computational Linguistics* 17 (1991) 21-48.

[10] T. Nomoto and Y. Nitta, Structuring raw discourse. In: D. Jones and H. Somers (eds.), *New Methods in Language Processing* (UCL Press, 1997) 288-298

[11] M. Phillips, Lexical structure of text. *Discourse Analysis Monographs*, vol. 12, (University of Birmingham Press, 1989)

[12] M. Pinto and C. Gálvez, Paradigms for abstracting systems, *Journal of Information Science* 25(5) (1999) 365-380.

[13] L. Polanyi, A formal model of the structure of discourse, *Journal of Pragmatics* 12 (1988).

[14] J.M. Ponte and W.B. Croft, Text segmentation by topic, *Proceedings of First European Conference on Research and Advanced Technology for Digital Libraries* (1997) 113-125.

[15] R. Rada and E. Bicknell, Ranking documents with a thesaurus, *Journal of American Society of Information Systems*, 40 (1989) 304-310.

[16] R. Rada, H. Mili, E. Bicknell, and M. Blettner, Development and application of a metric on semantic nets, *IEEE Transactions on Systems, Man, and Cybernetics* 19(1) (1989) 17-30.

[17] P. Resnik, Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language, *Journal of Artificial Intelligence Research* 11 (1999) 95-130.

[18] G. Salton, *Automatic Text Processing* (Addison-Wesley, 1989).

[19] G. Salton, A. Singhal, M. Mitra and C. Buckley, Automatic text structuring and summarization, *Information Processing and Management* 33 (1997) 193-207.

[20] S. Stoddard, *Text and Texture: Patterns of Cohesion. Advances in Discourse Processes, volume XL* (Ablex, 1991).

[21] R. Taraban and J.L. McClelland, Constituent attachment and thematic role assignment in sentence processing: Influences of content-based expectation, *Journal of Memory and Language* 27 (1988) 597-632.

[22] G. Whittemore, K. Ferrara and H. Brunner, Empirical study of predictive powers of simple attachment schemes for post-modifier prepositional phrases, *Proceedings of 28th Annual Meeting of the Association for Computational Linguistics* (1990) 23-30.

[23] Y. Yaari, Segmentation of expository texts by hierarchical agglomerative clustering, *Proceedings of the Interna-*

tional Conference of Recent Advances in Natural Language Processing, Tzigov Chark, Bulgaria (1995) 59-65.

[24] G. Youmans, A new tool for discourse analysis: the vocabulary-management profile, *Language* 67 (1991) 763-789.