Topic identification based on document coherence and spectral analysis

Joris D'hondt, Paul-Armand Verhaegen, Joris Vertommen, Dirk Cattrysse, Joost R. Duflou

Centre for Industrial Management, Katholieke Universiteit Leuven, Celestijnenlaan 300A bus 2422, 3001 Heverlee, Belgium

Abstract

In a world with vast information overload, well-optimized retrieval of relevant information has become increasingly important. Dividing large, multiple topic spanning documents into sets of coherent subdocuments facilitates the information retrieval process. This paper presents a novel technique to automatically subdivide a textual document into consistent components based on a coherence quantification function. This function is based on stem or term chains linking document entities, such as sentences or paragraphs, based on the reoccurrences of stems or terms. Applying this function on a document results in a coherence graph of the document linking its entities. Spectral graph partitioning techniques are used to divide this coherence graph into a number of subdocuments. A novel technique is introduced to obtain the most suitable number of subdocuments. These subdocuments are an aggregation of (not necessarily adjacent) entities. Performance tests are conducted in test environments based on standardized datasets to prove the algorithm's capabilities. The relevance of these techniques for information retrieval and text mining is discussed.

Key words: Topic identification, Spectral theory, Text mining

1. Introduction

In recent decades, the ever-increasing amount of digital texts has boosted the research for retrieval techniques to be able to deal with this vast amount of information. Clustering techniques and text categorization are a few examples of statistical techniques available for this purpose. Considering these full-text processes, the current computational limitations when dealing with large document datasets complicate the successful completion of these tasks. Digital texts or documents frequently describe more than one subject in their content. The ability to divide full text documents into components containing coherent parts or topics of these documents can help to bypass the computational limitations of these techniques. Information processing techniques on smaller documents decrease the required amount of memory and processing time, and therefore induce the research for supporting techniques such as text extraction and text summarization. These techniques reduce the size of documents and can be used as an input for further information processing. Topic-based segmentation has shown its usefulness for improved retrieval accuracy and retrieval of meaningful components of text, for document navigation and text summarization [2, 8].

The research introduced in this paper is focused on a novel technique that automatically subdivides a document in multiple topic-based components based on statistical and spectral properties of the text and the coherence graph of this text. The presented technique is able to discover non-contiguous connections, which differs from linear text segmentation in two main properties. Several of the existing techniques are applied to only identify topic boundaries between two adjacent document entities. No recombination of segments containing similar content is thus performed. The main difference between the presented technique and existing techniques is shown when these techniques are applied to the 21-paragraph Stargazers document [17], a well-known example in the context of topic identification. Subtopic frontiers were identified in this corpus by human judges in order to evaluate the TextTiling technique of Hearst. This subtopic structure is indicated below, together with the paragraph ranges:

- 1 → 3: Intro - the search for life in space
- 4 → 5: The moon's chemical composition
- 6 → 8: How early earth-moon proximity shaped the moon
- 9 → 2: How the moon helped life evolve on earth
- 13: Improbability of the earth-moon system
- 14 → 16: Binary/trinary star systems make life unlikely

- 17 → 18: The low probability of nonbinary/trinary systems
- 19 → 20: Properties of earth's sun that facilitate life
- 21: Summary

While subtopic algorithms try to identify these subtopic frontiers, the proposed technique tries to identify the general topics present in a document taking the overall content into account. Considering the Stargazers document, the presented technique identifies two general topics in this document: the first topic describes the earth-moon interaction, the second topic concerns the binary/trinary star systems hence indicating the difference between subtopic identification techniques.

The techniques described in this paper directly resulted in two applications, based on a multi-vector representation of a document [12]. A document vector in a vector space model [3] integrates all topics into one representation format which results in less accessible or even irretrievable information. The ability to divide full text documents into their components based on coherent parts or topics can help to bypass this retrieval issue. Constructing such multi-vector representation of a document aims to improve the retrieval performance of a search engine, because the information present in the content of the document is more accessible. The second developed application is the identification of near-duplicate documents based on the topics present in the content of the documents. As the content is more accessible, and the lexical chains are usable as a fingerprint of a document, interesting results are obtained [13].

This paper is structured in the following manner: in Section 2, a brief overview of related work in the domain top topic identification is presented. The subsequent section introduces the novel document segmentation technique based on topic identification. The four different steps in this topic identification and document segmentation process are explained in this section. The evaluation process is performed on four different test environments, based on two different scenarios. This procedure and the results are explained in Section 4. The paper closes with a section drawing conclusions from the obtained results.

2. Related work

As indicated in [10], the topic segmentation techniques can be divided in two categories: techniques respectively using statistical information extraction techniques and those exploiting lexical cohesion. This classification is however not strict. Several statistical methods adopt a format of lexical cohesion. Many topic identification algorithms assume that topically coherent subdocuments are related to text fragments exhibiting a homogeneous lexical distribution (i.e. the usage of words). In literature, several approaches can be identified, based on different descriptions of this distribution in a document.

The first category is based on lexical cohesion. Several approaches exist to measure this cohesion, such as stem of term repetitions, context vectors entity repetition, semantic similarity, word distance model and word frequency model. As mentioned in the previous section, text segments describing a similar content contain a similar vocabulary. The re-occurrence of specific terms can indicate the presence of a common topic. Lexical weighting is one of the most popular approaches in this type of topic identification [24, 17, 20, 31]. Lexical chains and the extended approach, the so-called weighted lexical links, are two techniques often used in a huge collection of identification algorithms. The topic unigram language model is the most frequently used technique [28]. Gathering the number of occurrences of each term for each topic leads to the posterior probability of a sequence of terms belonging to a certain topic. All terms obtain the same importance, i.e. terms not related to a topic are equally important as keywords. The so-called Cache model is based on a set of keywords automatically extracted for each topic. These words are the result of statistical distributions obtained from training corpora [5]. The TFIDF-classifier allows to represent each topic as a vector. These vectors contain the vocabulary specifically for their related topic. The similarity between a topic and a document represented in the Vector Space Model is calculated by the cosine similarity measure. The highest similarity indicates the topic of this document.

Most techniques based on this approach are linear topic segmentation algorithms. These algorithms place boundaries inside a text at positions where a topic shift is identified. This identification process is performed in a (fixed size) sliding window, examining lexical variations. The lexical variation often results in a drop of an employed similarity measure. As previously indicated, many algorithms can be described on this generic description. Popular examples are TextTiling [17], C99 [10], Dotplotting [29] or Segmenter [37]. The TextTiling technique segments texts into multiple entities (i.e. sequences of 3 to 5 sentences) or subtopics, the so-called 'tiles‘ using the cosine similarity between segments. A smoothed curve is calculated expressing the similarity between adjacent entities. Minima in this curve are considered as potential topic boundaries.

Other statistical approaches exist using global information of the text. Malioutov [22] presents a graph-theoretic framework. The text is converted into a weighted undirected graph in which the nodes represent the sentences and the edges quantify the relations. The text segmentation is performed by applying the normalized-cut criterion [30]. By using this criterion, the similarity within each partition is maximized and the dissimilarity across the partitions is minimized. The graph-based approach extends the local cohesion range of the sliding window by taking into account the long-range lexical cohesion and distribution in a text. The computational techniques for finding the optimal solution to the minimal cut objective are however difficult. The minimization of the normalized cut is NP-complete, but, due to the linearity constraint of this segmentation type, obtaining an exact solution is feasible [22].

All of the described algorithms rely on statistical properties of the text. The other category of techniques is based on Natural Language Processing techniques. Linguistic methods introduce a set of specific rules based on the corpus and use external semantic information such as thesauri and ontologies, possibly combined with one or more statistical methods[23]. This is the main drawback of this type of identification techniques: the results are dependent on the semantic resources available for a specific text [35] and therefore the setup is limited to the text. Hidden Markov Models and Neural Networks are used as part of the learning process in the technique of Amini [1]. A probabilistic sequence framework is proposed to estimate symbol or term sequences in a text. This framework should enable processing of more complex information retrieval and extraction tasks. Caillet proposed a machine learning technique based on term clustering [6]. The technique first discovers the so-called different concepts in a text, which are defined as sets of representative terms. The partitioning in coherent paragraphs is performed with the Maximum Likelihood clustering approach [19]. Passoneau and Litman [27] use decision trees in their algorithm to combine multiple linguistic features extracted from the document content. Other semantical techniques exist [11] that are able to recombine the segments according to their content.

A comprehensive overview of several, statistical and linguistic, topic identification techniques can be found in [7].

3. Document Segmentation

The underlying idea for the presented segmentation technique is common to the techniques proposed by Hearst [17] and Choi [10]: when a topic is described in a text, one can assume that a specific set of terms is used in the text fragment describing this topic. When a topic shift occurs, this set is substantially changed. Therefore re-occurrences of terms indicate a the presence of a certain topic, and topic boundaries can be identified. Identifying and quantifying the relationships between the so-called document entities (i.e. document parts, such as sentences or paragraphs) based on these terms enables the construction of a coherence graph. This graph represents the topical cohesion of the document, and is the main tool in the identification and segmentation process. Based on this graph, a matrix representation called the Laplacian matrix can be constructed with mathematically represents the connectivity of the graph. Using the spectral properties, the number and identification of the topically coherent components can be obtained. As these phases are matrix-based, the entire process is performed in an automatic manner.

The proposed segment identification process is thus composed of the following parts:

1. Construction of a coherence graph based on a linkage matrix
2. Construction of the normalized symmetric Laplacian of the coherence graph
3. Identification of the number of coherent components using the calculated Laplacian
4. Partitioning of the coherence graph using its spectral properties to obtain the components

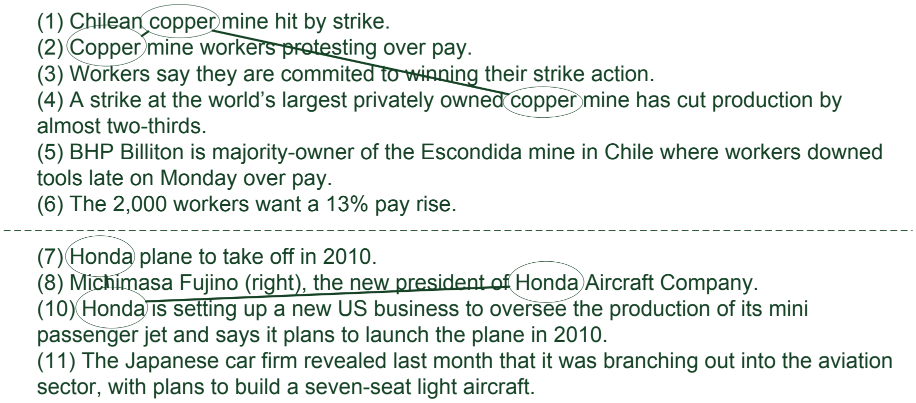
This last phase in the process is also a combination of two steps:

1. Reducing the feature space of the original document to a reduced subspace (see section 3.4)
2. Application of a cluster algorithm in this space to obtain the components

In the following sections, a profound overview of these phases is provided.

3.1. Coherence graph construction

The main idea of this document decomposition process is based on the presence of lexical chains (i.e. a chain of document entities based on the re-occurrence of a meaningful term in a document [4, 24]). The presence of these chains presents the lexical cohesive structure of the document [24]. The term and the identifiers of the document entities are stored as information of the chain. An example of such a lexical chain is given in Figure 1. In this figure, the sentences are taken as document entities. Their identifiers are shown between brackets. Several lexical chains are linking the sentences of a BBC article. The occurrences of the term 'copper' link the sentences 1,2 and 4; while the occurrences of term 'Honda' link the sentences 7, 8 and 10. These two lexical chains are only a subset of several possible lexical chains in this small example. Considering all chains present in this example, two larger components can be identified. The first component is the set of sentences from 1 to 6, the second component contains the set of sentences from 7 to 11. In the figure, these two components are separated by a horizontal line.



(1) Chilean copper mine hit by strike.
(2) Copper mine workers protesting over pay.
(3) Workers say they are commited to winning their strike action.
(4) A strike at the world's largest privately owned copper mine has cut production by almost two-thirds.
(5) BHP Billiton is majority-owner of the Escondida mine in Chile where workers downed tools late on Monday over pay.
(6) The 2,000 workers want a 13% pay rise.

(7) Honda plane to take off in 2010.
(8) Michimasa Fujino (right), the new president of Honda Aircraft Company.
(10) Honda is setting up a new US business to oversee the production of its mini passenger jet and says it plans to launch the plane in 2010.
(11) The Japanese car firm revealed last month that it was branching out into the aviation sector, with plans to build a seven-seat light aircraft.

Figure 1: Examples of stem chains

In a lexical chain not all the document entities are related in a similar manner. The first difference is based on the notion of distance. The notion of distance between document entities in this context is defined as the number of entities they span. In Figure 1 the distance between the first and third occurrence of 'copper' is 4 (sentence 1 to 4). Intuitively, the nearer the related document entities are, the more chance there is that they are related to a similar content. However, it remains possible that another distant set of entities is related with the same stem. These entities are also describing a similar content, and both groups will form one longer chain. However, due to the distance, not all pairwise relationships in this chain will be equally strong.

A second difference between the entities is the informative value of a word. Not all words in a document are equally important [25]. An obvious example are stopwords. These carry almost no content, and therefore can be considered as noise when creating stem chains. Longer chains are also preferred over shorter ones. This reasoning is based on the importance of mid-frequency words originating from the Zipf-curve [25]. Considering the possibility that a term occurs multiple times within a document entity, the longest chains are most likely defined by the most frequently occurring terms in the document. This requires an optimal filtering of non-informative terms. Shorter chains can be created based e.g. on typing mistakes or the presence of numerical values.

The previous assumptions are translated into a quantification formula. The linkage $w_{i,j,C}$ between document entities i and j based on stem chain C is defined as

$$w_{i,j,C} = w_C \cdot Nb_C \cdot \frac{\|D\|}{1 + (C_i - C_j)} \tag{1}$$

w_C is the weight of the term defining the chain C in the document, such as raw or normalized frequency [19]. Nb_C is the number of occurrences in the term chain C, $\|D\|$ is the length of the document expressed in number of document entities, and $C_i - C_j$ is the distance between entities i and j in the text. In a complete document, multiple term chains can be identified and quantified. The overall linkage $w_{i,j}$ between document entities i and j is based on the aggregation of all linkage values for the term chains that connect both entities:

$$w_{i,j} = \sum_C \cdot w_{i,j,C} \tag{2}$$

To illustrate this process, consider the following example in Table 1. This document contains 5 document entities (indicated as rows), connected by 6 term chains shown in the columns.

Applying this process between all pairs of document entities and for all identified chains, a square linkage matrix W can be obtained containing all the linkage values. The linkage matrix of the example presented in Table 1 is shown in the following matrix

Table 1: Example of a document containing 5 document entities and 6 lexical chains

Entity	**Term 1**	**Term 2**	**Term 3**	**Term 4**	**Term 5**	**Term 6**
A	1	1	0	1	1	0
B	1	1	0	0	0	1
C	0	0	1	1	0	0
D	0	0	1	0	0	0
E	0	1	0	0	1	1

$$W = \begin{bmatrix} 0 & 8.05 & 1.83 & 0 & 1.12 \\ 8.05 & 0 & 0 & 0 & 2.55 \\ 1.83 & 0 & 0 & 3.22 & 1 \\ 0 & 0 & 3.22 & 0 & 0 \\ 1.12 & 2.55 & 0 & 0 & 0 \end{bmatrix}$$

Based on this matrix, a graph structure can be composed to represent the quantified relationships between the document entities. This undirected, weighted coherence graph is defined as:

$$G = (V, E) \tag{3}$$

comprising a set V of vertices, representing the document entities, together with a set E of edges representing the linkage values. This means that each edge between two vertices v_i and v_j carries a non-negative weight $w_{i,j}$.

In Figure 2, the coherence graph of the example presented in Table 1 is shown. The dots in circular form are the different nodes representing the document entities, the lines represent the coherence between the related pair of document entities. The weight values are omitted for reasons of clarity.

A real example of a graph structure is given in Figure 3. The final graph partitioning result is also indicated in this figure. After applying the process which explained in the next sections, three topics can be identified. The cuts are indicated with a full line.

3.2. Graph Laplacian construction

The previously obtained graph provides a (graphical) representation of the topical coherence of the document. By employing a graph partitioning technique on this graph, the weakest connections are removed and thus the different topics are obtained. To obtain a graph partitioning of the coherence

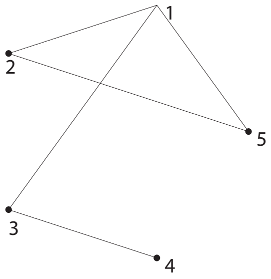


Figure 2: Coherence graph of a document containing 5 document entities.

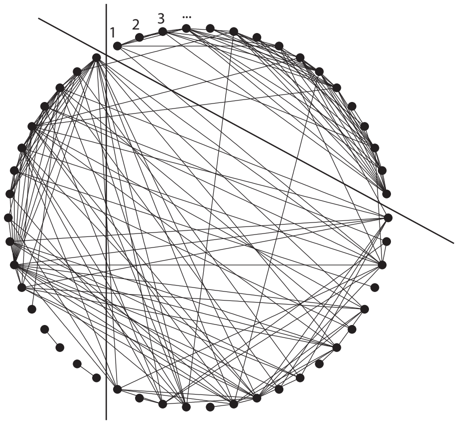


Figure 3: Coherence graph of a document containing 50 document entities.

graph, spectral graph partitioning techniques can be employed [36]. The main tool for this spectral graph partitioning are the so-called Laplacian matrices. With every graph such Laplacian matrix can be constructed, which is a mathematical representation of the graph. In literature multiple definitions of a Laplacian matrix can be found [36]. The symmetric normalized Laplacian matrix is used for reasons of stability and consistency [36]. This square matrix, based on the number of document entities, is defined as

$$L = I - D^{-\frac{1}{2}} \cdot W \cdot D^{-\frac{1}{2}} \tag{4}$$

I and D respectively denote the identity matrix and the degree matrix of the graph. W represents the previously obtained linkage matrix. In this context, the following definition for a degree matrix is used:

Definition A degree matrix D of a weighted, undirected graph is a diagonal matrix restricted to the following conditions:

$$D = \begin{cases} d_{i,j} = \sum_{k=1}^{n} w_{i,k} & \text{if i = j} \\ 0 & \text{otherwise} \end{cases}$$

Considering the example in Table 1, the obtained degree matrix of the graph is presented in the following matrix:

$$D = \begin{bmatrix} 10.99 & 0 & 0 & 0 & 0 \\ 0 & 10.60 & 0 & 0 & 0 \\ 0 & 0 & 5.05 & 0 & 0 \\ 0 & 0 & 0 & 3.22 & 0 \\ 0 & 0 & 0 & 0 & 3.67 \end{bmatrix}$$

Using this matrix in Equation 4, the following Laplacian matrix is obtained:

$$L = \begin{bmatrix} 1 & -0.75 & -0.25 & 0 & -0.18 \\ -0.75 & 1 & 0 & 0 & -0.41 \\ -0.25 & 0 & 1 & -0.80 & 0 \\ 0 & 0 & -0.80 & 1 & 0 \\ -0.18 & -0.41 & 0 & 0 & 1 \end{bmatrix}$$

3.3. Number of topics identification

As stated in [36], the second smallest eigenvalue of the linkage matrix explains a significant amount of variance or structure present in the coherence graph[36]. The eigenvector related to this eigenvalue thus indicates the connectivity between the different document entities. If the linkage matrix is sorted in the order proposed by this eigenvector, the reordered matrix provides indications of the different document entities that are related together according to quantification of the different lexical chains.

Reordering the linkage matrix of the example shown in Table 1 according to the order proposed by this eigenvector, the following matrix is obtained:

$$W = \begin{bmatrix} 1 & -0.41 & -0.74 & 0 & 0 \\ -0.41 & 1 & -0.18 & 0 & 0 \\ -0.74 & -0.18 & 1 & -0.25 & 0 \\ 0 & 0 & -0.25 & 1 & -0.80 \\ 0 & 0 & 0 & -0.80 & 1 \end{bmatrix}$$

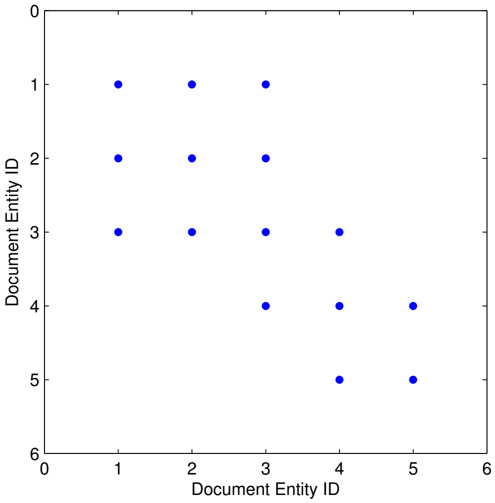


Figure 4: Matrix representation of the coherence graph of the example document

If this matrix is graphically represented, the Figure 4 is obtained. The dots in this figure indicate existing relationships between the document entities. A more complex example of this reordering process is shown in Figure 5.

The reordered matrix thus contains dense segments along the diagonal indicating coherent subgraphs in the coherence graph. The ideal situation is that the segments have a full square form. These squares indicate the presence of complete subgraphs.

The identification process of the number of topics based on the coherence graph is based on this property. The number of components can be determined by identifying the relevant dense segments along the diagonal. In an optimal configuration these blocks are dense squares, as indicated in this formula:

$$L = \begin{bmatrix} L_1 & 0 & 0 \\ 0 & L_2 & 0 \\ 0 & 0 & L_3 \end{bmatrix}$$

with L_i a full dense matrix.

For the further identification process, the notion of a square segment along the diagonal is introduced. This shape is defined to have its upper left and lower right corner on the diagonal, and its size is defined as the number

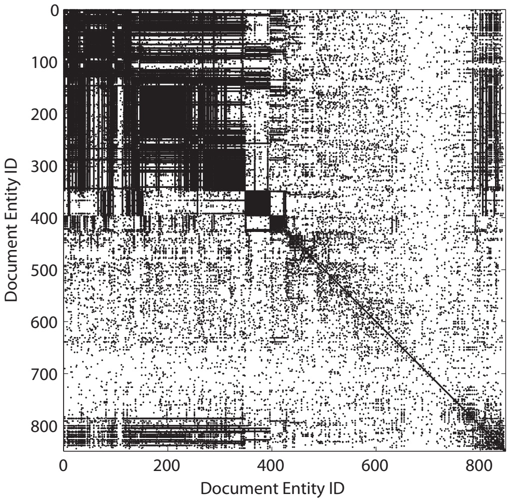


Figure 5: Matrix representation of the coherence graph

of document entities it spans.

The automatic identification process is composed of two steps. The first step is to extract a number of square segments possibly indicating coherent components based on a single threshold parameter. The second step is to retain those segments covering a substantial number of relationships.

In the first step, the following algorithm 1 is used to retrieve dense segments. For every row, the sizes of possible segments are identified. The size of a segment is defined by T, the number of adjacent non-zero values in a row of the matrix representation. A threshold value THR stipulates how many adjacent zero values, indicated as blank spaces in Figure 5, can be taken into account when delineating the segment.

The result of this first step in the process is a list of possible square segments. The segments can overlap with, or be included in, other segments. The number of segments involved in every non-overlapping list, is a possible result of the number of components problem.

The second step in the identification process uses the notion of coverage ratio of a square segment. This ratio R_i of square segment i is defined as

$$R_i = \frac{S_{i,incl}}{S_{i,incl} + S_{i,excl}} \tag{5}$$

$S_{i,incl}$ is the number of relationships that are included in the identified

```
Input: Linkage matrix $L$ sorted in the order of the second largest
       eigenvector
while $T <$ threshold $THR$ # of non-adjacent relations do
    foreach Document entity $i$ do
        j=i; if $L(i,j) > 0$ then
            $count++; j++$
        end
        else
            THR–;
        end
        if $THR = 0$ then
            exit;
        end
        else
            i++;
        end
    end
    foreach Document entity $i$ do
        Create a square segment with count as size, and upper left
          corner positioned at point $i$
    end
    foreach Document entity $i$ do
        Calculate the density of the square. The density is defined as
          the number of related document entities over the total
          possible number of relationships. If the density of this square
          is above a predefined threshold, retain this square.
    end
end
Output: Enumeration of square segments
```

Algorithm 1: The Square Segment Identification (SSI) algorithm

square segment. $S_{i,excl}$ is the number of relationships that are excluded from the square. In Figure 6 the coverage percentage of the indicated segment is 0.667. Due to the symmetry of the linkage matrix, the analysis of half the matrix is sufficient. The upper triangle of the square segment, starting at element 2 with a size of 3, contains 6 non-zero values. The total number of non-zero values covered by rows of the square segment is nine, indicated

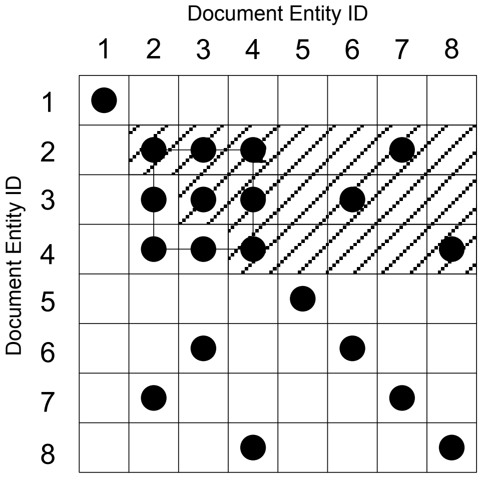


Figure 6: Example of the coverage ratio calculation. The ratio for the hatched region in this example is 0.667.

in Figure 6 as hatched cells. For every identified square segment, this coverage percentage can be calculated. As a result, for every non-overlapping list of square segments, the total coverage ratio can be calculated as the average sum of the coverage of these segments. This score is situated in the range]0, 1]. The higher this fraction, the higher the chance is that the corresponding number of components will represent the actual number that can be distinguished in the document. The reasoning is the following: the larger the group of of entities that is covered by a non-overlapping list of square segments, the better this number of square segments will indicate the coherent structure present in the document. These square segments are defined by the variance explained by the second largest eigenvector, and therefore are clustered based on common properties. A value of one for the coverage indicates a full coverage of the entities by the related number of square segments, a near 0-result indicates the opposite.

The algorithm identifies 2 topics with high probability (ratio score of 1) for the example of Table 1. The ratio score for the second possibility, which is 1 topic, is 0.6. For the example shown in Figure 5, a number of square segments can be identified. The SSI-algorithm returns 4 as the most likely number of components, followed by 6, 2 and 3. Less likely are 16 and 19.

3.4. Spectral partitioning of the graph

As already introduced in the previous section, the actual partitioning of the coherence graph is based upon the 'minimal cut approach'. This approach results in the partitioning of the graph by removing those edges that have the lowest sum of coherence quantification in order to obtain the previously obtained number of subgraphs. Given the graph $G = (V, E)$, minimizing a cut of a graph is mathematically similar to the identification of a non-trivial vector x that minimizes the following function

$$F(x) = \sum_{(i,j)\in V} e_{i,j}(x_i - x_j) \tag{6}$$

In this objective function x is a bi-partition (A, B) vector with $x_i = 1$ if $i \in A$ and $x_j = -1$ if $j \in B$. $e_{i,j} \in E$ is the weight of the relation between vertices i and j. This objective function can be reformulated as

$$x^T \cdot L \cdot x \tag{7}$$

with x the bi-partition vector and L the Laplacian matrix. This formula describes the relationship between the minimal cut of the graph and its spectral properties. The solution of Equation 7 is the so-called Fiedler vector [16], which is used in the previous section. The eigenvalue is also called the 'algebraic connectivity' of the graph, and is greater than 0 if and only if the graph is connected. A graph is called connected if every pair of distinct vertices in the graph can be connected through some sequential path of vertices.

The partitioning process is a two-phase partitioning or clustering process. Starting from the representation of the Laplacian, the spectral or Jordan normalized decomposition can be calculated. Selecting the k eigenvectors related to the largest eigenvalues of the Laplacian matrix, with k equal to the identified number of components, a k-dimensional representation in an inner product space is obtained [32]. This mapping of the original data into a new metric space expresses the alignment or coherence of the graph, rather than its structure based upon the original similarity. The second phase is the actual clustering step. Any clustering algorithm can be employed to obtain a partitioning of the coherence graph. In this research, the algorithm of Ng, Jordan and Weiss [26] is applied to partition the coherence graph in multiple subgraphs. The clustering algorithm used in this phase is the agglomerative hierarchical clustering algorithm using average linkage [19]. This type of algorithm is known to perform well in a document clustering environment

[14]. An overview of this process is given in [36]. The resulting subgraphs can be converted into the required coherent components by reordering the document entities.

Applying the segmentation process on the example shown in Table 1 results in the following ordering: the first, second and fifth document entity are clustered together, while the third and fourth document entity form the second topic.

4. Validation

The validation process has two main objectives, in accordance with the two phases in the overall identification process:

1. validation of the technique concerning the number of identified topics
2. validation of the quality of the identified topics

The first test environments described in Section 4.3 aim to validate the novel technique of identifying the number of components, as well as to compare the retrieval performance of the extracted components considering the known structure of the documents. The second validation part compares the quality of the presented technique to two well-known subtopic identification techniques:

- Choi (C99)[10]
- TextTiling [17]

These techniques were chosen based upon their consistent performance under various circumstances [15, 33].

This second validation step is merely to present the resemblances and the differences between the topic and the selected subtopic identification technique.

4.1. Datasets

For the experimental validation of these techniques, the following datasets were used to create the required test environments:

- Reuters RCV1 document dataset [21]
- OHSUMED medline dataset [18]

The Reuters RCV1 test collection is one of the most widely used collections for text categorization, containing Reuters articles assigned to various categories on different levels of detail.

4.1.1. Reuters

Six categories were selected from the Reuters RCV1 dataset with the following labels:

- GCRIM: Craw, law enforcement
- GDEF: Defense
- GDIP: International relations
- GDIS: Disasters and accidents
- GENT: Arts, culture, entertainment
- GSPO: Sports

This selection is based upon the size and the number of documents present in each category.

4.1.2. Ohsumed

From the OHSUMED test collection only the documents were withheld rated as definitely relevant with one of the queries, as stated in the OHSUMED descriptions. This resulted in a document collection containing 101 topics (five queries returned no documents), with in total 1985 documents. All the documents in this dataset, are truncated to a maximum of 250 words.

4.2. Preprocessing

The following steps were automatically performed on each document to obtain a suitable input format:

- punctuations and other marking symbols were removed
- stopword removal

4.3. Validation process

Two test scenarios were constructed using the previously described standardized datasets OHSUMED and Reuters resulting in four experimental setups. Both scenarios create artificial documents by concatenating original documents. The goal of these scenarios is to identify the topics present in the artificial documents.

The first test scenario is the sequential adding of a random number of randomly selected original documents. In Figure 7, the left document is a graphical representation of this construction process. This artificial document is composed of 5 document parts, originating from three documents. The different document parts of these documents are added sequentially. All document entities of the selected original documents are concatenated to form the artificial documents. The document entities in this test environment are paragraphs delimited by blank lines in the original document. As the complexity of the complete process is quadratic in the number of sentences, each paragraph is limited to ten sentences so larger paragraphs were split. Every newly constructed document contains a random number of categories of the applied dataset. The number of selected documents varies between two to ten documents. Duplicates and multiple documents from the same directory are allowed. For each standardized dataset a collection of these artificial documents is created.

The second test scenario differs in the order in which the document entities are concatenated. In this test scenario the document entities are concatenated in a random manner to effectively distribute the different topics present in the document. This idea is shown in Figure 7 as the right document. This document is also composed of 5 document parts originating from three documents. The document parts are however added randomly to the document, in order to have recurring topics in the content of the artificial documents. The other conditions are also valid for this test scenario.

These two test scenarios result in four different test sets, two for each standardized dataset. Each generated dataset contained 1000 artificial documents. The results of the topic identification techniques on these test sets are discussed in Section 5.

4.4. Validation measure

Comparing two types of topic identification techniques, identifying topics at different levels of refinement, can pose a problem. Since the focus of the proposed set of techniques is a reordering the document entities according to their spectral properties, an adapted F-measure based validation is used. The original F-measure is a popular measure for evaluating a clustering result [9]. This measure for a cluster i is a weighted harmonic mean of precision and recall of cluster i, and is mathematically defined as

$$F_i = \frac{2 \cdot recall_i \cdot precision_i}{recall_i + precision_i} \tag{8}$$

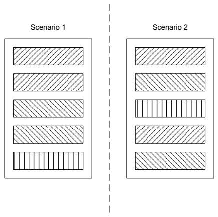


Figure 7: Schematic representation of the two different test scenarios

To calculate the $precision_i$ and $recall_i$, each cluster must be assigned to a known topic label. The adapted F-measure assumes a cluster to represent a topic. The notion of topic in this validation measure is related to the category the document entity originates from. To obtain this measure the following scenario is used:

- Find the largest group of document entities on the same topic within a document
- Label this group of entities as the group entities covering that topic
- Mark this group as labeled, and the topic label as assigned
- Repeat with the remaining groups and topics that are left
- Check that every group is covering a different topic. If not, the group containing the smaller group of document entities on the same topic is set to be relabeled to a different topic.
- Repeat until all groups are covering a different topic.

Based upon this process, a F-measure value can be obtained for every processed artificial document [9].

5. Results

In this section results are presented of the application to the three topic identification techniques for the four test environments. For each artificial document of each of these test environments, the initial composition or topics are known. Therefore comparisons can be made based on the identification of the topic boundaries. These boundaries are obtained from the three topic identification techniques, thus indicating the functional differences between the two existing techniques, and the newly proposed topic identification technique. This difference is situated on the level of detail on which the presented techniques identify topics, and the ability to identify non-continuous topics. The clustering technique used in the second phase of the topic identification process was the unsupervised hierarchical agglomerative clustering method using average linkage [19]. For every test environment the results are summarized as boxplots [34] in which the five represented numerical values are defined as:

- Smallest observation
- First quartile Q1 (25 %)
- Median Q2
- Third quartile Q 3(75 %)
- Largest observation

Observations that are identified as outliers are also indicated. An observation is considered as a weak outlier if it is located between the bounding quartiles and 1,5 times the interquartile range $Q3 - Q1$. If it is located outside this range, it is considered as a strong outlier.

5.1. Sequential test environments

Figure 8 presents the boxplot of the obtained F-measures for the first Ohsumed test set. The indicator values determining this boxplot are given in Table 2. A difference in average of 8 % can be noted between the proposed technique and the technique of Choi. The range of obtained F-measures for the proposed technique is significantly smaller compared to the Choi and TextTiling techniques. However, the positions of the boxplots of Choi and the proposed technique are positioned similar along the range of F-measures, indicating a slightly better performance for the proposed technique. This indicates that the identification techniques of Choi and the proposed techniques do not differ significantly for this sequential test environment. Finally,

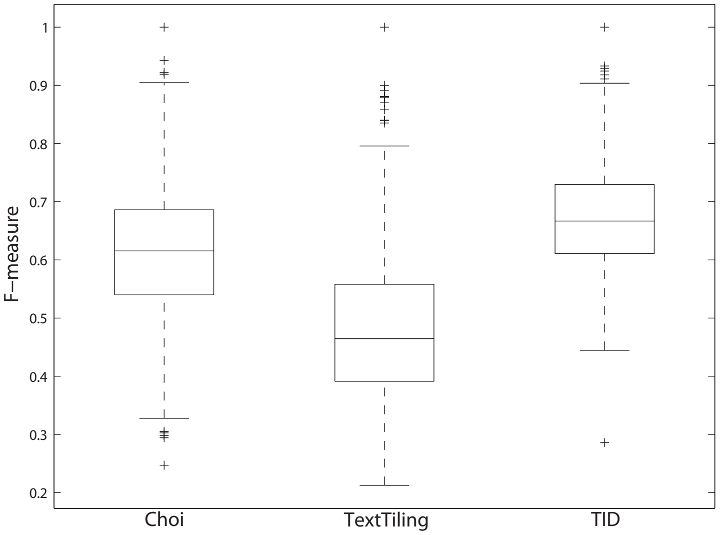


Figure 8: Boxplot of the sequential Ohsumed test environment

Table 2: Quality results of the sequential Ohsumed test environment

Dataset	Choi	TextTiling	TID
First quartile	0.5398	0.3913	0.6106
Median	0.6154	0.4644	0.6667
Third quartile	0.6892	0.5581	0.7296
Average	0.6122	0.4766	0.6774

no strong outlier are identified among the obtained F-measures for each of the test techniques.

The results for the second sequential test environment are summarized in Figure 9. The related figures determining this boxplot are given in Table 3. Similar conclusions can be drawn as in the first sequential test environment considering the position of the boxplot of the proposed technique compared to TextTiling and Choi. No significant difference in position of the boxplot is notable between the three techniques. The differences in averages are considerably larger: the average of the proposed technique is 22,71 % higher than Choi and 10.9 % than TextTiling. Also the range of the boxplot of the proposed technique is the smallest of all three obtained boxplots. Two observations were considered as strong outliers, one for TextTiling and the proposed technique each.

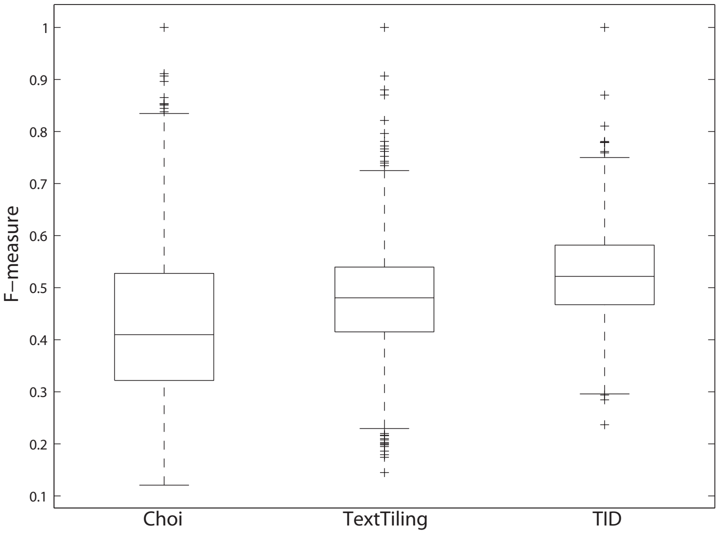


Figure 9: Boxplot of the sequential Reuters test environment

Table 3: Quality results of the sequential test set

Dataset	**Choi**	**TextTiling**	**TID**
First quartile	0.3218	0.4152	0.4672
Median	0.4096	0.4806	0.5217
Third quartile	0.5275	0.5396	0.5817
Average	0.4327	0.4784	0.5310

Based on the results of these 2000 artificial documents, sequentially composed of the Reuters and Ohsumed datasets, the conclusion can be drawn that the proposed technique performs similar or better compared to two well-known topic identification techniques in the context of identifying the topic boundaries.

5.2. Randomized test environment

The results of the second test scenario, in which artificial documents are composed in random manner, are presented in this section. Figure 10 presents the boxplot of the obtained F-measures for the first randomized Ohsumed test set with the related scores displayed in Table 4.

The first significant difference between the three boxplots, is the position of the boxplot of the proposed technique. The boxplot range of 0.4672 to

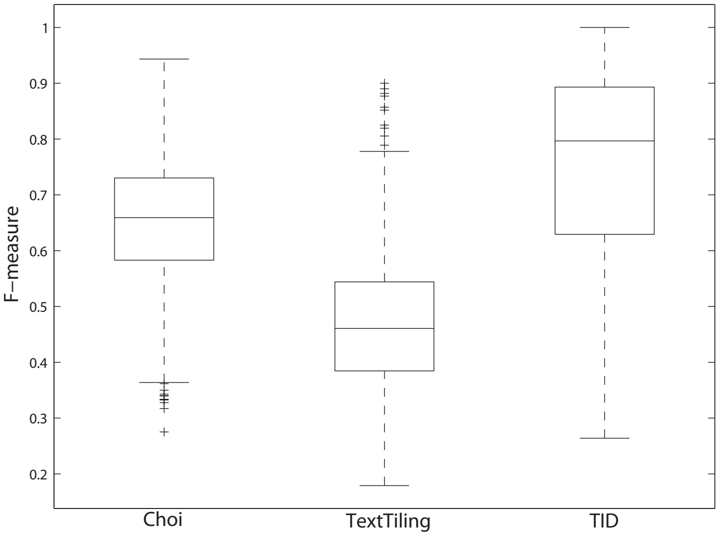


Figure 10: Boxplot of the Ohsumed test environment

Table 4: Quality results of the random Ohsumed test set

Dataset	Choi	TextTiling	TID
First quartile	0.5829	0.3846	0.6292
Median	0.6592	0.4606	0.7966
Third quartile	0.7304	0.5439	0.8932
Average	0.6553	0.4694	0.7457

0.5817 is significantly higher than the other two boxplots. Also the complete ranges of all three boxplots are notable larger than in the first test environments, the largest range being for the presented technique. This is mainly due to the nature of the technique. Since some documents, chosen randomly out the datasets without any content knowledge, mostly contain only numerical information, the proposed technique fails to identify the expected structure. Choi and TextTiling are less error prone in this situation compared to the proposed technique, indicated by the range of the boxplots. No strong outliers are identified during the identification processes.

The results of the last test environment, composed of Reuters articles, are summarized in Figure 11 with the related scores of the boxplot shown in Table 5. These results are in line with the results of the previous random test environment: a clear distinction can be noted between the position of

Table 5: Quality results of the random Reuters test set

Dataset	Choi	TextTiling	TID
First quartile	0.3391	0.3925	0.5942
Median	0.4292	0.4667	0.7049
Third quartile	0.5457	0.5287	0.7842
Average	0.4516	0.4642	0.6712

the boxplot of the presented technique and the position of the second highest boxplot of Choi.

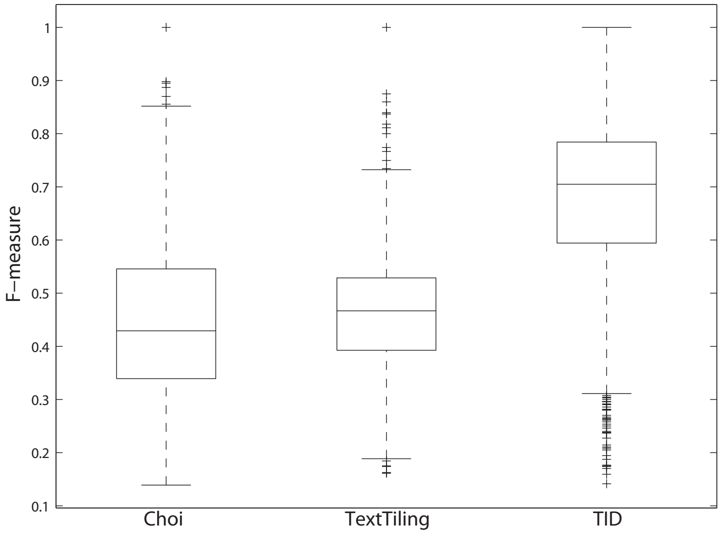


Figure 11: Boxplot of the Reuters test environment

6. Conclusions

A novel technique was introduced to automatically identify the topics present in a document, based on the presence of lexical chains. Since every topic can be related to a specific set of words, the coherence of a document can be quantified using these lexical chains and the proposed similarity measure. The application of spectral graph partitioning techniques on the coherence graph transforms the original input documents to a new metric space. This transformation has two benefits in the topic identification process. First, the required number of topics can be identified based on the reordering of the Laplacian matrix of the graph. Second, a clustering technique can be applied on this new metric space in order to obtain the required number of components, each representing a topic.

This topic identification technique differs from existing approaches because these current techniques identify subtopic document parts. Two large experiments based on standardized datasets were performed to validate these techniques. The results indicate that the techniques resulted in similar to better subtopic identification results in a sequential test scenario, whereas in a randomized test scenario the proposed technique outperforms the two other subtopic identification technique. This last result indicates that the application of these graph techniques enables the identification of non-contiguous topics in a document.

References

[1] M. Amini, H. Zaragoza, and P. Gallinari. Learning for sequence extraction tasks. In *Content-Based Multimedia Information Access*, pages 476–490, 2000.

[2] R. Angheluta, R. D. Busser, and M.-F. Moens. The use of topic segmentation for automatic summarization. In *In Workshop on Text Summarization in Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization*, pages 11–12, 2002.

[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison Wesley, May 1999.

[4] R. Barzilay and M. Elhadad. Using lexical chains for text summarization. In *In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 10–17, 1997.

[5] B. Bigi, R. de Mori, M. El-Béze, and T. Spriet. A fuzzy decision strategy for topic identification and dynamic selection of language models. *Signal Processing*, 80(6):1085–1097, 2000.

[6] M. Caillet, J.-F. Pessiot, M.-R. Amini, and P. Gallinari. Unsupervised learning with term clustering for thematic segmentation of texts. In *Proceedings of Seventh Conference on Rercherche d'Information Assistee par Ordinateur*, pages 648–656, 2004.

[7] Y. Chali. Topic detection of unrestricted texts: Approaches and evaluations. *Applied Artificial Intelligence*, 19(2):119–135, 2005.

[8] L. Chen, J. Zeng, and N. Tokuda. A "stereo" document representation for textual information retrieval. *J. Am. Soc. Inf. Sci. Technol.*, 57(6):768–774, 2006.

[9] T. Y. Chen, F.-C. Kuo, and R. G. Merkel. On the statistical properties of the f-measure. In *QSIC*, pages 146–153, 2004.

[10] F. Y. Y. Choi. Advances in domain independent linear text segmentation. In *In Proceedings of NAACL*, pages 26–33, 2000.

[11] C. Clifton, R. Cooley, and J. Rennie. Topcat: Data mining for topic identification in a text corpus. *IEEE Transactions on Knowledge and Data Engineering*, 16(8):949–964, 2004.

[12] J. D'hondt. *Clustering Techniques in Knowledge Management: Advances and Applications.* PhD thesis, Katholieke Universiteit Leuven, Leuven, Belgium, February 2011.

[13] J. D'hondt, P. Verhaegen, J. Vertommen, D. Cattrysse, and J. Duflou. Near-duplicate detection based on text coherence quantification. In *In Proceedings of the 10th European Conference on Knowledge Management*, pages 238–246, 2009.

[14] J. D'hondt, J. Vertommen, P.-A. Verhaegen, D. Cattrysse, and J. R. Duflou. Pairwise-adaptive dissimilarity measure for document clustering. *Inf. Sci.*, 180:2341–2358, June 2010.

[15] G. Dias, E. Alves, and J. G. P. Lopes. Topic segmentation algorithms for text summarization and passage retrieval: an exhaustive evaluation. In *AAAI'07: Proceedings of the 22nd national conference on Artificial intelligence*, pages 1334–1339. AAAI Press, 2007.

[16] M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23:298–305, 1973.

[17] M. A. Hearst. Texttiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

[18] W. Hersh, C. Buckley, T. J. Leone, and D. Hickam. Ohsumed: an interactive retrieval evaluation and new large test collection for research. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 192–201, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[19] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, September 1999.

[20] M. G. Kathleen and K. Mckeown. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 562–569, 2003.

[21] D. Lewis, Y. Yang, T. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research, 2004.

[22] I. Malioutov and R. Barzilay. Minimum cut model for spoken lecture segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 25–32. Association for Computational Linguistics, 2006.

[23] M.-F. Moens and R. D. Busser. Generic topic segmentation of document texts. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 418–419, New York, NY, USA, 2001. ACM.

[24] J. Morris and G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48, 1991.

[25] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46, 2005.

[26] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press, 2001.

[27] R. J. Passonneau and D. J. Litman. Discourse segmentation by human and automated means. *Computational Linguistics*, 23:103–139, 1997.

[28] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281, New York, NY, USA, 1998. ACM.

[29] J. C. Reynar. Statistical models for topic segmentation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 357–364, Morristown, NJ, USA, 1999. Association for Computational Linguistics.

[30] J. Shi and J. Malik. Normalized cuts and image segmentation. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, page 731, Washington, DC, USA, 1997. IEEE Computer Society.

[31] L. Sitbon and P. Bellot. Topic segmentation using weighted lexical links (wll). In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 737–738, New York, NY, USA, 2007. ACM.

[32] D. B. Skillicorn. *Understanding Complex Datasets: Data Mining with Matrix Decompositions*. Chapman & Hall,CRC, 2007.

[33] N. Stokes. Spoken and written news story segmentation using lexical chains. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 49–54, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[34] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[35] M. Utiyama and H. Isahara. A statistical model for domain-independent text segmentation. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 491–498, 2001.

[36] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, December 2007.

[37] M. yen Kan, J. L. Klavans, and K. R. Mckeown. Linear segmentation and segment significance. In *In Proceedings of the 6th International Workshop on Very Large Corpora*, pages 197–205, 1998.