Automatic Construction of News Hypertext

Theodore Dalamagas

15 Sep 1997
Abstract

Hypertext information retrieval systems combine hypertext and information retrieval capabilities by providing retrieval techniques that include direct search as well as navigation and browsing. The incorporation of a mechanism for the automatic construction of hypertext into an IR system is quite important in order to exploit the advantages of the integration of hypertext with information retrieval.

News hypertext deals with articles from newspaper archives. Formal ways to capture the temporal aspects that characterize the newspaper domain, and which are often ignored by IR systems, are presented. The suggested framework is conceptualized with the notions of stories and threads, which are substories within a story. Threads are identified by applying clustering techniques to articles’ segments that correspond to subtopics within the main topic of an article and then linking the segments which belong to the same cluster.

The evaluation of such an approach to the automatic construction of hypertext is finally presented, in terms of its usability and the structural quality of the resulting hypertext.
Acknowledgements

I am grateful to my supervisor Mark D. Dunlop for his outstanding guidance and valuable advise, who took a keen interest for my work all over the year, both in the implementation part and the writing of this thesis.

I acknowledge the support of Marcos Theophylactou for the endless debates that we had for information retrieval and the exchanging of so many useful ideas for our theses.

I would like also to thank all the AIS students for the collaboration that we had all over the year. I single out Fotis Lazarinis for the numerous helpful discussions that we had.

Many thanks to Marcos, Sibylle, Katja, Simone and Gabriela, whose company made my residence so pleasant in our mixed student flat, as well as to all my new friends that I met here in Glasgow.

Finally, my warmest thanks to my parents and my brother for their support and enthusiasm during my studies in Glasgow.

Theodore Dalamagas
September 1997
Glasgow
...to Anna Maria
Contents

1 Introduction 1
  1.1 Thesis motivation ........................................... 1
  1.2 Thesis aim .................................................. 2
  1.3 Outline ..................................................... 2

2 Information retrieval and hypertext 4
  2.1 Overview .................................................... 4
  2.2 Information retrieval ........................................ 4
     2.2.1 Automatic text analysis ................................. 4
     2.2.2 Retrieval strategies ..................................... 9
     2.2.3 Evaluation ............................................... 11
  2.3 Hypertext ................................................... 12
     2.3.1 Basic concepts .......................................... 12
     2.3.2 Graph-based representation of hypertext ............... 14
     2.3.3 Towards information retrieval ........................... 17
  2.4 Automatic construction of hypertext ........................ 18
  2.5 Conclusion .................................................. 19

3 News hypertext 21
  3.1 Overview .................................................... 21
  3.2 The newspaper paradigm ...................................... 21
     3.2.1 Retrieval requirements for newspaper domains ........... 22
  3.3 Formal aspects of news hypertext ............................ 23
  3.4 Conclusion .................................................. 26
4 NHS: News Hypertext System
   4.1 Overview .......................................................... 27
   4.2 Introduction ....................................................... 27
   4.3 Retrieval engine ................................................... 29
   4.4 Article decomposition ............................................ 31
      4.4.1 Segment detection based on paragraph relationship maps 31
      4.4.2 Implementation issues ....................................... 32
   4.5 Thread construction .............................................. 34
      4.5.1 Clustering methods .......................................... 35
      4.5.2 Implementation issues ....................................... 36
   4.6 NHS prototype .................................................... 40
   4.7 Conclusion .......................................................... 41

5 Evaluation
   5.1 Overview .......................................................... 42
   5.2 Introduction ....................................................... 42
   5.3 Segment evaluation ................................................. 43
   5.4 Thread evaluation .................................................. 44
      5.4.1 Hypertext metrics .............................................. 44
      5.4.2 User judgements .............................................. 47
   5.5 Overall evaluation ............................................... 51
   5.6 Conclusion .......................................................... 54

6 Further work
   6.1 Introduction ....................................................... 55
   6.2 Implementation level .............................................. 55
   6.3 Evaluation process ................................................. 57

7 Conclusions ........................................................... 58

Bibliography .............................................................. 60

A The NHS prototype ..................................................... 65
B Sample of the evaluation instructions
List of Figures

2.1 Zip’s law ................................................................. 7
2.2 Matrix representation of graphs ................................. 16
2.3 Compactness and stratum metrics for various hypertext structures 17
3.1 Temporal layout of story evolution .......................... 25
3.2 Object model: threads and stories .............................. 26
4.1 Architecture of the NHS system ............................... 28
4.2 Simplification of paragraph relationship map for segment detection ... 32
4.3 Primitive schemes of distant linked paragraphs .............. 33
4.4 Extending chains of adjacent paragraphs ...................... 34
4.5 News hypertext ......................................................... 35
4.6 Maximum Spanning Tree (MST) detection and single link clustering 38
4.7 The NHS system: data flow .................................. 41
5.1 Expected form of news hypertext ............................ 46
5.2 Overall evaluation results for the NHS system .............. 52
A.1 The main page of the NHS system ......................... 65
A.2 Retrieved articles from the NHS system ..................... 66
A.3 An example of an article which is part of a thread ........... 67
A.4 An example of an article which is part of a thread ........... 68
A.5 The temporal layout that the NHS system provides .......... 69
A.6 The story evolution of a thread ............................... 70
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Automatic links</td>
<td>13</td>
</tr>
<tr>
<td>3.1</td>
<td>An example of a thread</td>
<td>24</td>
</tr>
<tr>
<td>4.1</td>
<td>Refinements of a single link clustering process</td>
<td>39</td>
</tr>
<tr>
<td>5.1</td>
<td>Segment evaluation results from user experiments</td>
<td>44</td>
</tr>
<tr>
<td>5.2</td>
<td>Example of a decomposed article</td>
<td>45</td>
</tr>
<tr>
<td>5.3</td>
<td>News stories</td>
<td>46</td>
</tr>
<tr>
<td>5.4</td>
<td>Compactness and stratum for the resulting hypertext of the five stories</td>
<td>47</td>
</tr>
<tr>
<td>5.5</td>
<td>Evaluation of thread quality</td>
<td>47</td>
</tr>
<tr>
<td>5.6</td>
<td>An example of a false thread</td>
<td>48</td>
</tr>
<tr>
<td>5.7</td>
<td>An example of a summarizing thread</td>
<td>49</td>
</tr>
<tr>
<td>5.8</td>
<td>Examples of threads</td>
<td>50</td>
</tr>
<tr>
<td>5.9</td>
<td>An example of a non-relevant thread</td>
<td>53</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Thesis motivation

Newspaper domain provides a large base of information that can be handled by an information retrieval (IR) system. Traditional IR systems treat the newspaper domain as a flat collection of documents by providing the user with a set of articles relevant to a query. However, they do not take into account two basic aspects of newspaper domain which are not found in other domains:

- Newspaper articles tend to be clustered by topics.
- Newspaper articles often refer to topics close but slightly different from a main topic.
- Temporal information, like the sequence of the events, is quite important.

Hypertext is helpful when dealing with the above aspects. Structural links can be used in order to create a hierarchy of articles based on content. The user can browse the hierarchy according to her information need. Links between related terms can be also put to capture related notions in different articles. A recent proposal of integrating hypertext and information retrieval especially for the newspaper domain can be found in (Golovchinsky, 1997), in which the VOIR system is described. In this work, the link construction is based on feedback from users. Query terms that distinguish well among articles become candidate nodes for hypertext.

However, despite the usage of structural links and term links, the temporal aspect is still ignored. Furthermore, term linking does not provide a sufficient way to relate similar notions in different articles. This makes the information search difficult to perform and time consuming.

Thus, a different way of handling news related documents is needed, giving also importance to the notion of story evolution. By presenting the sequence of the events to the user...
and not only a flat set of retrieved articles, the understanding of the information provided is quite easier and, thus, the satisfaction of her information need faster.

Automatically constructed hypertext which captures the notion of story evolution and conveys information extracted by the sequence of events should be incorporated into information retrieval systems that deal with the domain of newspaper archives.

1.2 Thesis aim

This thesis presents a methodology for the automatic construction of links for news hypertext which is tailored to the domain of newspaper archives. This methodology needs to be formally conceptualized, implemented and tested for its effectiveness.

Thus, the aim of this thesis can be summarized as follows:

- It will set the requirements for information retrieval on the domain of newspaper archives.
- It will provide a way to model formally these requirements so as to take advantage of the hypertext paradigm.
- It will prove the applicability of the above model by describing the implementation of an integrated Hypertext Information Retrieval system (HIR), called News Hypertext System (NHS), that automatically creates hypertext.
- It will present methods to evaluate the retrieval effectiveness of the NHS system.

1.3 Outline

This thesis is organized as follows:

Chapter 2 presents an overview of information retrieval and hypertext. Automatic indexing and weighting techniques as well as various retrieval strategies are discussed. Matrix representations of graph structures and their usage for analysing the structure of a hypertext are described. Finally, various aspects that refer to the integration of information retrieval and hypertext, among them the automatic construction of hypertext, are presented.

Chapter 3 describes the newspaper paradigm in information retrieval. It sets the requirements of the newspaper domain, which generally are different from those of other domains. The notions of stories and threads, which are substories within a story, are suggested to deal with the newspaper domain.

Chapter 4 presents the framework for the implementation of the NHS system, which is based on the model suggested in chapter 3. The basic retrieval engine is briefly discussed.
A method for article decomposition into segments is presented, followed by a clustering technique to group related segments. Finally, a WWW based prototype of the NHS system is described and its architecture is analysed.

Chapter 5 deals with the evaluation methods which were used to determine the effectiveness of the NHS system. The evaluation process was performed in three parts: segment evaluation, thread evaluation and overall evaluation. User tests as well as hypertext topological metrics were used.

Finally, chapter 6 summarizes issues that need to be further considered, both in the implementation level of the NHS system and the evaluation process.
Chapter 2

Information retrieval and hypertext

2.1 Overview

This chapter presents an overview of information retrieval (IR) and hypertext. A typical IR system provides a way to search for documents in response to a query. Documents and queries are represented by a set of descriptors, which are produced using automatic text analysis techniques. Automatic indexing and weighting are discussed in this chapter as the main parts of the automatic text analysis for information retrieval. Another important aspect of an IR system is the way which a document is characterized as relevant or non-relevant. Retrieval strategies, including the vector spacing retrieval and the probabilistic retrieval, are presented and evaluation methods are briefly discussed.

Hypertext is a non-linear way of managing textual documents. Matrices that represent its graph-based structure provide a simple way to make an analysis of this structure. Such kind of matrices are described and two metrics that deal with the complexity and the linearity of hypertext are analyzed. Moreover, the relationship between hypertext and information retrieval is examined in terms of an integrated hypertext information retrieval system.

Automatic construction of hypertext is one of the important aspects for designing and building hypertext information retrieval systems. A review of this field is finally presented.

2.2 Information retrieval

2.2.1 Automatic text analysis

One major problem in information retrieval systems is the representation of the content of the documents. Text analysis should be applied in order to assign to each document a
set of descriptors capable of representing its content. The assignment of the descriptors should fulfił three related purposes (Salton and McGill, 1983; Keen, 1977):

1. “To allow the location of items dealing with topics of interest to the user.”

2. “To relate items to each other, and thus relate the topic areas, by identifying distinct items dealing with similar, or related, topic areas.”

3. “To predict the relevance of individual information items to specific information requirements through the use of index terms with well-defined scope and meaning.”

The process of descriptor assignment can be carried out automatically, using a computer system, or manually, using subject experts. The former is known as automatic indexing (as opposed to manual indexing) and is the one that will be discussed here. In addition to the descriptor assignment, the text analysis should provide a way of measuring the importance of each descriptor for purposes of content identification, which is known as weighting. A brief description of both processes of text analysis for information retrieval, automatic indexing and weighting, based on statistical methods\(^1\), follows.

**Automatic indexing**

The automatic indexing process starts with the identification of each individual word from the document as a possible descriptor.

Following the identification of words, the stop-words need to be eliminated. Stop-words are high frequency words that by all means are poor descriptors when taking out of context (articles, adverbs, prepositions etc). A portion of such a set of words can be found in (van Rijsbergen, 1979). Stop-words not only are useless for content identification but also comprise about 50 percent of the document text.

The next step is the removal of word suffixes and prefixes so that each word is reduced to its stem. The process is called stemming and is used to improve retrieval effectiveness and reduce the size of indexing files. Stemming will cause the words ‘walking’ and ‘walked’ to be reduced to the common stem ‘walk’.

Stemming generally either improves the retrieval effectiveness or has no effect at all. For an overview of many experiments with stemming algorithms as well as a discussion of several implementations, including the well-known Porter stemming algorithm (Porter, 1980), one can refer to (Frakes, 1992). Stemming algorithms often produce errors. For example, Porter’s algorithm causes the words ‘public’ and ‘publication’ to be reduced to the common stem ‘public’, although the two words should be distinguished. The assumption is that the proportion of such kind of errors does not have a real effect on the retrieval performance.

---

\(^1\) Linguistic methods based on natural language processing had mixed results when were incorporated into information retrieval systems. See (Sparck Jones, 1990) for an extreme suggestion.
Each stem that has been detected from a document can be one of its descriptors, which are known as index terms. However, index terms are usually more sophisticated than just stems which have been extracted from documents. Dictionaries are used, so that concepts related to the index terms are also included in the set of the index terms in the hope of broadening their interpretation. Broad interpretation can be also achieved by using a thesaurus, which provides a grouping of the terms (keywords) used in a given topic area into categories. The process is called automatic keyword classification and can be exploited either by replacing each term of the descriptors set by the name of the category it belongs to, or by replacing each term of the descriptors set by all the keywords of the category it belongs to.

After word stems are generated as possible index terms, the weighting process follows in order to identify those index terms that have great importance for content identification.

**Weighting**

The weighting process assigns a weight to each index term according to its importance for content identification. Most weighting methods are based on the observation that the frequency of occurrence of a word in a text is related to its importance for content representation. If the individual words of a document are presented in decreasing order of their frequency of occurrence in that document, then the Zipf’s constant rank-frequency law (Zipf, 1949) is typically observed: In a document, the frequency of a word multiplied by its rank order is equal to the frequency of another word multiplied by its rank order (see figure 2.1). Zipf’s law is explained from the fact that people usually prefer repeating already used words to creating new ones.

Luhn specified two cut-offs for the Zipf’s law graph (Luhn, 1958). Words below the lower cut-off were characterized as “common”, without having significant contribution to content identification. Words above the upper cut-off were characterized as “rare”, without again having significant contribution to content identification. Words between the two cut-offs were considered to have significant content identification power. Lyhn’s ideas are also visualized in figure 2.1.

Based on the above remarks, Lyhn suggested that only words with medium collection frequency should be used as index terms. The collection frequency for a word \( w \) is determined by summing the frequencies of \( w \) across all documents of the collection.

However, these ideas are too strict in the context of information retrieval. Words with low collection frequency can be helpful in having many relevant documents in the set of the retrieved ones, while words with high collection frequency can be helpful in having most of the relevant documents retrieved. What really characterizes a word as a useful index term is the fulfilment of the following requirements (Salton and McGill, 1983):

\[ \text{rank 1 is assigned to the most frequent word} \]

\[ \text{neither high nor low} \]
It must be related to the information content of its document so as to make it retrievable when it is wanted, thus having a great proportion of the relevant documents retrieved. The proportion of the relevant documents that are retrieved is known as recall.

- It must distinguish its document from the remainder in order to prevent the retrieval of all items, wanted or not, thus having a great proportion of the retrieved documents to be relevant. The proportion of the retrieved documents that are relevant is known as precision.

Terms with high frequency of occurrence in a document seem to be useful to the first requirement. This suggests the usage of a term frequency \((tf)\) factor as the first part of the weighting scheme.

Terms with low frequency of occurrence in the whole collection seem to be useful to the second requirement. This suggests the usage of an inverse document frequency \((idf)\) factor as the second part of the weighting scheme.

Using the product of the term frequency \(tf_{ij}\) and the inverse document frequency \(idf_j\) for a term \(j\) of a document \(i\), one can obtain a good measure of the importance of that term for content identification of that document by using the following weight \(w_{ij}\) for that term (Salton and Yang, 1973):

\[
    w_{ij} = tf_{ij} \cdot idf_j
\]  

(2.1)

The simplest form of the \(tf\) component is the binary one: \(tf\) is equal to 1 for terms present in a document and 0 to the others. A set of other possible forms of the \(tf\) component can
be found in (Salton and Buckley, 1988). Two typical forms follow:

\[ t f_{ij} = rt f_{ij} \]  

(2.2)

where \( rt f_{ij} \) is the raw term frequency (number of times the term \( j \) occurs in the document \( i \)), and

\[ t f_{ij} = 0.5 + 0.5 \cdot \frac{rt f_{ij}}{\max(rt f_{i})} \]  

(2.3)

where \( \max(rt f_{i}) \) is the maximum raw term frequency for any term of the document \( i \). The latter form is known as augment term frequency.

A typical \( idf \) component may be estimated as (Sparck Jones, 1972):

\[ idf_j = \log \left( \frac{N}{f_j} \right) \]  

(2.4)

where \( f_j \) is the total number of occurrences of term \( j \) in the collection and \( N \) the number of the documents in the collection.

In addition to the \( tf \) and \( idf \), a normalization factor appears useful, especially in collections of documents with widely varying lengths. Longer documents tend to have a better chance to be retrieved as relevant, although all relevant documents, no matter long or short they are, should be treated as equally important. The normalized \( tf \cdot idf \) weighting can be defined as follows (Salton and Buckley, 1988):

\[ w_{ij} = \frac{t f_{ij} \cdot idf_j}{\sqrt{\sum_j (w_{ij}^2)}} \]  

(2.5)

The \( tf \cdot idf \) weighting model does not exhibit substantial theoretical properties, as opposed to the probabilistic weighting. According to the latter, an appropriate weight \( w_j \) for a term \( j \) is given by the following expression (Robertson and Sparck Jones, 1976):

\[ w_j = \frac{r/(R - r)}{(n - r)/[N - n - (R - r)]]} \]  

(2.6)

where \( R \) is the number of relevant documents, \( r \) is the number of relevant documents that contain the term \( j \), \( N \) is the number of all documents in the collection and \( n \) is the number of documents that contain the term \( j \). The above expression is known as relevance weight and defines the importance of a term by using the proportion of relevant documents in which the term occurs divided by the proportion of non-relevant documents in which the term occurs. At first, relevance weights seem to be strongly dependent on relevance information that is difficult to be available a priori. However, methods have been proposed for using relevance weights without (Croft and Harper, 1979) or with little relevance information (Robertson and Walker, 1997).
Assuming term independence (the terms occur independently of each other) and binary term scheme, it has been shown (see (Croft and Harper, 1979; van Rijsbergen, 1979)) that the probabilistic relevance weight becomes quite similar to the \( tf \cdot idf \) weighting. A brief discussion about the probabilistic model of information retrieval will be presented in the following section, together with other retrieval models.

### 2.2.2 Retrieval strategies

In case of an information retrieval system, documents are retrieved which are considered relevant to a user’s query. All retrieval strategies are based on a comparison between the query and the documents, which identifies possible relevant documents to the specific query.

#### Vector space retrieval

Vector space retrieval (Salton and Yang, 1975; Salton and McGill, 1983) considers a document space consisting of documents. The three-dimensional space is extended to \( u \) dimensions when \( u \) index terms are present. For each document \( i \), \( u \)-dimensional document vectors \( \mathbf{D}_i \) are constructed from a set of \( u \) index terms \( t_1, t_2, \ldots, t_u \):

\[
\mathbf{D}_i = (d_{i1}, d_{i2}, \ldots, d_{iu})
\]  
(2.7)

where \( d_{ij} \) is the weight that is assigned to term \( j \) for document \( i \).

Similarly, a \( u \)-dimensional vector \( \mathbf{Q} \) is constructed for the query that a user inserts:

\[
\mathbf{Q} = (q_1, q_2, \ldots, q_u)
\]  
(2.8)

where \( q_j \) is the weight that is assigned to term \( j \) for query \( Q \).

Salton suggests the raw term frequency for document weighting and the augment term frequency for query weighting, both combined with the \( idf \), for most of the retrieval cases (Salton and Buckley, 1988).

Using the vector representations (2.7) and (2.8), similarity values are computed for each document-query pair:

\[
Sim(\mathbf{D}_i, \mathbf{Q}) = \frac{\sum_{j=1}^{u} (d_{ij} \cdot q_j)}{\sqrt{\sum_{j=1}^{u} d_{ij}^2 \cdot \sum_{j=1}^{u} q_j^2}}
\]  
(2.9)

Documents that their similarity value with the query is above a predefined threshold are considered to be relevant to that query. As a result, the final output of a vector-space based IR system is a set of documents, usually ranked in descending order of similarity value, related to the user’s query.
Probabilistic retrieval

Probabilistic retrieval takes into account the relevance properties of the documents. According to the binary independence retrieval model (Robertson and Sparck Jones, 1976; Croft and Harper, 1979; van Rijsbergen, 1979), which is the simplest one in probabilistic retrieval, each document (query) is represented by a \( n \)-dimensional binary vector \( \mathbf{x} \) (\( \mathbf{r} \)):

\[
\mathbf{x} = (x_1, x_2, \ldots, x_u) \tag{2.10}
\]

\[
\mathbf{r} = (r_1, r_2, \ldots, r_u) \tag{2.11}
\]

where \( x_j \) (\( r_j \)) indicates the absence or presence of the \( j \)th term in the document (query), by being equal to 0 or 1 respectively.

A document is relevant for a particular query if the probability of the document being relevant, given the document vector \( \mathbf{x} \), is greater than the probability of the document being non-relevant:

\[
P(\text{Relevant}|\mathbf{x}) > P(\text{Non-relevant}|\mathbf{x}) \tag{2.12}
\]

From the above decision rule, the following matching function is derived, from which retrieval status values \( g \) are computed for each document-query pair:

\[
g(\mathbf{x}, \mathbf{r}) = \sum_{j=1}^{u} \left( r_j \cdot x_j \cdot \log\frac{p_j(1 - q_j)}{(1 - p_j)q_j} \right) + C \tag{2.13}
\]

where \( p_j \) is the probability that if the document is relevant, the index term \( j \) will be present, whereas \( q_j \) is the probability that if the document is non-relevant, the index term \( j \) will be present. \( C \) is constant for a given query and does not affect the ranking of the documents. Documents are ranked in descending order of the retrieval status values.

One way to estimate the probabilities \( p_j \) and \( q_j \) is by having an initial search based on other retrieval strategies and using the top few documents as relevant, or by applying user relevance feedback (see subsequent paragraphs). However, the above probabilities can be also estimated without (Fuhr and Huther, 1989) or with little relevance information (Robertson and Walker, 1997).

Assuming that all \( p_j \) are the same and that \( q_j \) is estimated by \( n_j/N \), where \( n_j \) is the number of documents in which the term \( j \) occurs and \( N \) is the size of the collection, then the probabilistic matching function becomes quite similar to the vector space matching function (2.9) with binary weighting used (Croft and Harper, 1979).

Other retrieval strategies and techniques

Other retrieval strategies include the boolean retrieval model and the cluster-based model. In the former, each document is related to a set of keywords and each query has the form
of a boolean expression with and, or and not operators. The retrieved documents are those which contain index terms in the combination specified by the query. In cluster-based models, documents are grouped into clusters. Clusters provide another mechanism for matches between query terms and document clusters, in addition to the conventional matches between query terms and document terms.

Relevance feedback is a well-known technique used in many retrieval strategies in order to increase the retrieval effectiveness. It can be achieved either by reweighting of query terms based on the distribution of these terms in the set of relevant and non-relevant documents retrieved in response to this query, or by changing the actual terms in the query. Users judge the relevance of the retrieved documents after an initial search has been performed. For a detailed description of relevance feedback techniques, one can refer to (Harman, 1992).

2.2.3 Evaluation

Information retrieval evaluation has mainly focused on retrieval effectiveness and has been based on the usage of document collections, queries and relevance judgements given by users expert in the subject of the retrieval queries, although these judgements are often subjective. Documents retrieved in response to a query are judged by these users as being relevant or not.

The most common measures of retrieval effectiveness is recall ($RE$) and precision ($PR$), which have been introduced implicitly during the description of weighting methods in section 2.2.1. Recall is the proportion of the relevant documents that are retrieved in response to a query and precision is the proportion of the retrieved documents that are relevant.

For the rank position $i$ of each relevant document in response to a query, precision $PR_i$ and recall $RE_i$ values are calculated:

$$PR_i = \frac{r_i}{n_i}, \quad RE_i = \frac{r_i}{R}$$

where $R$ is the number of relevant documents, $r_i$ is the number of relevant documents returned at that point of rank position $i$ and $n_i$ is the number of documents returned at that point of rank position $i$. The $PR$, $RE$ values for each query are interpolated and then the average values are calculated for all the queries in order to have a set of precision values at recall points of 0.1, 0.2, … 1, from which graphs are constructed.

The TREC conference is a benchmark exercise of IR systems of many IR groups. Each group uses the same set of queries in order to retrieve documents from the same collection. Their top-ranked documents are sent back to the TREC, pooled and judged for relevance. Precision-recall graphs are constructed and averages of precision values are calculated for a standard number of retrieved documents.

A combined measure of recall and precision, which reflects the importance that an exper-
CHAPTER 2. INFORMATION RETRIEVAL AND HYPERTEXT

The experimenter gives to each one of them, has been suggested by (van Rijsbergen, 1979):

\[ E = 1 - \frac{(1 + b^2) \cdot P \cdot R}{b^2 \cdot P + R} \]  

(2.15)

For example, having \( b = 0.5 \) the experimenter is twice as interested in precision as recall.

For a detailed analysis of various IR evaluation methods one can refer to (Salton and McGill, 1983; van Rijsbergen, 1979). User evaluation aspects for information retrieval systems can be found in (Dunlop, 1996).

2.3 Hypertext

2.3.1 Basic concepts

Hypertext is defined as a non-linear, non-sequential way of managing textual documents. In hypertext, each document is considered to be segmented in fragments. Each fragment is represented by a node. Fragments can be related to each other by links. Thus, hypertext can be visualized as a network of nodes that are connected by links.

Links establish logical relationships between the nodes that they connect and form paths in the hypertext which a user can browse. A node can be an index one if it has many links to other nodes. A reference node is the converse of the former: it has many links towards it. The process of constructing hypertext is known as authoring and is performed with the help of hypertext authoring tools.

Initially, hypertext was used to refer to textual documents. However, nowadays documents do not consist only of text data but they include also graphics, audio, video, animation etc. The term hypermedia is sometimes used, instead of ‘hypertext’, to deal with such kind of data.

Link taxonomy

In hypertext, links connect nodes on the basis of a relationship between them. A link type is a description of such a relationship. Parunak makes an extensive classification of link types that are useful in hypermedia (van Dyke Parunak, 1991). Allan presents an amalgam of link types that Parunak suggests, by separating them into three classes according to the difficulty of their identification and creation (Allan, 1996):

- Pattern-matching links: This class represents links that deal with the layout or the logical structure of a document. A good example is links between chapters, links from a bibliographic citation to the cited work, links from words to their definitions etc. Pattern-matching links are easy to be identified and created automatically from a hypertext system.
Automatic links

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>revision link</td>
</tr>
<tr>
<td>summary link</td>
</tr>
<tr>
<td>equivalence link</td>
</tr>
<tr>
<td>comparison link</td>
</tr>
<tr>
<td>aggregate link</td>
</tr>
</tbody>
</table>

Table 2.1: Automatic links

- **Manual links**: This class represents links that deal with relationships that are on the level of natural language understanding. A good example is links between the important parts of a debate that is described in a sequence of documents. Manual links are quite difficult to be identified and created automatically from a hypertext system without human intervention.

- **Automatic links**: This class represents links that deal with revision, summary, equivalence, comparison, and aggregation. Most of these links can be characterized as content links, because they connect parts of documents (or just documents) with similar content. Details about this class are presented in table 2.1. Automatic links can be identified and created automatically from a hypertext system with fair success.

**Hypertext usability**

Taking into consideration the various link types that were presented, many advantages can be indicated by incorporating hypertext in computer systems. The operational advantages of hypertext are summarized as follows (Conklin, 1987):

- Unstructured information can be organized hierarchically.
- Text segments can be grouped together in many ways, allowing the same document to serve multiple functions.
- Documents can be presented in many ways: by a table of contents, by global or local views, etc.
- References are handled easily.
- Information is modular and consistent.

On the other hand, two major problems arise using hypertext (Conklin, 1987):
CHAPTER 2. INFORMATION RETRIEVAL AND HYPERTEXT

- **Disorientation**
  Hypertext provides more freedom in the way that the user has access to information. Disorientation comes when the browsing of the hypertext is complex because of the great number of links. In this occasion, the user cannot easily understand how to use the hypertext in order to find the information that she wants.

- **Cognitive overload**
  Hypertext is constructed during the authoring process under a specific *cognitive model* that the user has to understand and learn in order to use hypertext properly. A user may face cognitive overload in case she is not able to reach a sufficient level of understanding this cognitive model.

2.3.2 Graph-based representation of hypertext

Although there is a clear relationship between hypertext structures and *graphs*, the analysis of the former on the basis of formal graph-theoretic concepts hadn’t started until late ’80s (see for instance the Neptune system (Delisle and Schwartz, 1987)). Since then, many graph-based models have been suggested for a formal analysis of the functions dealing with objects (nodes) and relations (link) in a hypertext or hypermedia structure. Lucarella for example suggests a graph-based object model for general hypermedia structures based on *hypermedia schemas* which consist of *classes*, *class attributes* and *property relationships* (Lucarella and Zanzi, 1996).

However, the basic *directed-graph model* remains fundamental, especially for hypertext structures which are much simpler than hypermedia structures. Having a graph \( G \) with a set of vertices \( V \) and edges \( E \), each node \( n \) of a hypertext corresponds to a vertex \( v \in V \) and each link to an edge \( e \in E \). Graphs can be easily represented by matrices. Three well-known graph matrix representations follow (for a brief discussion of these representation methods see also (Furner et al., 1996)).

**Matrix representation of graphs**

1. For any graph \( G \) with \( n \) vertices \( v_j \in V \), \( j = 1 \ldots n \), and \( p \) edges \( e \in E \), the \( n \times n \) adjacent matrix \( A(G) \), with elements \( \alpha_{ij} \), is defined:

\[
\alpha_{ij} = \begin{cases} 
1, & \text{if } v_j \text{ is adjacent to } v_i, \ i \neq j \\
0, & \text{if } v_j \text{ is not adjacent to } v_i, \ i \neq j \\
\text{NULL}, & \text{if } i = j 
\end{cases} \quad (2.16)
\]

The sum of all values in the \( i \)th row gives the *out-degree* \( o_d_i \), of vertex \( v_i \). Similarly, the sum of all values in the \( j \)th column gives the *in-degree* \( i_d_j \), of vertex \( v_j \). In non-directed graphs, the out-degree of each vertex is equal to its in-degree.
2. For any graph $G$ with $n$ vertices $v_j \in V$, $j = 1 \ldots n$, and $p$ edges $e \in E$, the $n \times n$ distance matrix $A(G)$, with elements $\delta_{ij}$, is defined:

$$
\delta_{ij} = \begin{cases} 
  nL, & \text{nL is the number of links of any shortest} \\
  & \text{path from } v_i \text{ to } v_j, \ i \neq j \text{ and } v_i \text{ is connected to } v_j \\
  0, & \text{if } v_i \text{ is not connected to } v_j, \ i \neq j \\
  NULL, & \text{if } i = j 
\end{cases}
$$

(2.17)

The sum of all values in the $i$th row gives the status $i, s_i$, of vertex $v_i$. Similarly, the sum of all values in the $j$th column gives the contrastatus $j, c_j$, of vertex $v_j$. In non-directed graphs, the status of each vertex is equal to its contrastatus.

3. For any graph $G$ with $n$ vertices $v_j \in V$, $j = 1 \ldots n$, and $p$ edges $e \in E$, the $n \times n$ converted distance matrix $A_c(G)$, with elements $\delta_{ij}^c$, is defined:

$$
\delta_{ij}^c = \begin{cases} 
  nL, & \text{nL is the number of links of any shortest} \\
  & \text{path from } v_i \text{ to } v_j, \ i \neq j \text{ and } v_i \text{ is connected to } v_j \\
  n, & \text{if } v_i \text{ is not connected to } v_j, \ i \neq j \\
  NULL, & \text{if } i = j 
\end{cases}
$$

(2.18)

The sum of all values in the $i$th row gives the converted out-distance $i, cod_i$, of vertex $v_i$. Similarly, the sum of all values in the $j$th column gives the converted in-distance $j, cid_j$, of vertex $v_j$. In connected non-directed graphs, the converted distance matrices are identical to the distance ones.

In figure 2.2, examples of the above graph matrix representations are depicted.

Matrices that represent graphs provide the source for identifying various properties of graph structures by calculating metrics based on their data elements. Thus, using graphs to represent hypertext and having their corresponding matrices, one can be able to identify various characteristics of the hypertext structure. Such kind of metrics are presented in the following paragraphs.

**Hypertext metrics**

Botafogo suggests *compactness* and *stratum*, two metrics that capture notions of complexity and linear ordering in a hypertext structure (Botafogo et al., 1992). Their values vary between 0 and 1.

Compactness deals with the complexity of the hypertext. High compactness ($\cong 1$) indicates that there are many links among the nodes of the hypertext, as opposed to low compactness ($\cong 0$). A completely connected hypertext has compactness of value 1, whereas
CHAPTER 2. INFORMATION RETRIEVAL AND HYPERTEXT

Figure 2.2: Matrix representation of graphs

<table>
<thead>
<tr>
<th>GRAPH G</th>
<th>ADJACENT MATRIX(G)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>i,j</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>IDj</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DISTANCE MATRIX(G)</th>
<th>CONVERTED DISTANCE MATRIX(G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>i,j</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>1</td>
<td>-1 1 2 1 6</td>
</tr>
<tr>
<td>2</td>
<td>0 -1 0 0 1</td>
</tr>
<tr>
<td>3</td>
<td>0 0 -0 0 0</td>
</tr>
<tr>
<td>4</td>
<td>0 1 1 -0 2</td>
</tr>
<tr>
<td>5</td>
<td>0 2 2 1 -5</td>
</tr>
<tr>
<td>Cj</td>
<td>0 4 6 3 1 14</td>
</tr>
</tbody>
</table>

Figure 2.2: Matrix representation of graphs

A completely unconnected one has compactness of value 0. Compactness is defined as:

$$C_p = \frac{(n^3 - n^2) - \sum_i \sum_j \delta_{ij}^c}{n^3 - 2n^2 - n}$$  \hspace{1cm} (2.19)

where $n$ the number of the nodes of the hypertext and $\delta_{ij}^c$ the elements of its converted distance matrix.

Stratum provides a way to examine the linear ordering of hypertext. High stratum ($\cong 1$) indicates a linear structure hypertext structure, whereas low stratum ($\cong 0$) indicates a cyclical one. A completely linear hypertext has stratum of value 1, whereas a completely cyclic one has stratum of value 0. Stratum is defined as:

$$St = \begin{cases} \frac{\sum_i (s_i - c_i)}{n}, & n \text{ is even} \\ \frac{\sum_i (s_i - c_i)}{n^2 - n}, & n \text{ is odd} \end{cases}$$  \hspace{1cm} (2.20)

Some examples of compactness and stratum metrics are presented in figure 2.3.
2.3.3 Towards information retrieval

Hypertext is quite important when it comes to the area of information retrieval. Traditional IR system are based on the assumption that the documents to be retrieved and presented are linear. Hypertext changes this linearity and helps the user to dynamically construct her own information path, for example by (Agosti, 1996)

- linking a document to a term (or the opposite)
- connecting two related terms
- connecting two related documents or parts of them

_Hypertext information retrieval (HIR)_ systems combine hypertext and information retrieval capabilities by providing retrieval techniques that include direct search (access by content) as well as navigation and browsing (access using associations).

New retrieval models are needed to support not only direct search but also navigation and browsing. In (Dunlop and van Rijsbergen, 1994), free text retrieval and hypermedia are
combined using contextual information for the retrieval of non-textual documents. Extensions of the probabilistic retrieval model to include hypertext information have also been suggested (Croft and Turtle, 1993). Finally, an integrated model for hypermedia and information retrieval has been proposed by (Chiaramella and Kheirbek, 1996). Following this model, a HIR system deals with content knowledge and structural knowledge. The former describes the semantic content of data, whereas the latter deals with information that is conveyed by the links between data. Content and structural knowledge are represented using conceptual graphs (Sowa, 1984). The derived knowledge base together with documents, queries and a matching function compose the four basic components of an integrated hypermedia IR system. For an implementation of this model one can refer to (Berrut et al., 1995).

Several other aspects should be taken into account for the design of an effective HIR system (Agosti, 1996):

- Since the incorporation of hypertext capabilities in an IR system increases its complexity, new user modelling and interface issues arise.
- Techniques to evaluate effectively HIR systems should be developed.
- The notion of active hypertext is another necessary aspect. Ways of attaching a level of importance to links must be implemented so that different link relevances are provided.

Finally, automatic construction of hypertext, the automatic transformation of a flat collection into a hypertext, is quite important, especially if the collection consists of a huge number of documents. The difficulty for accomplishing such a transformation varies according to the type of links that are needed (see section 2.3.1). The following section presents briefly various information retrieval techniques for the automatic construction of hypertext.

### 2.4 Automatic construction of hypertext

Methods for the automatic construction of hypertext document collections have been considered by researchers as an important part of the hypertext authoring process (Agosti, 1996). Following the link taxonomy that has been described in section 2.3.1, automatic hypertext creation can be achieved relatively easy in case of structural links, which represent the layout or the logical structure of a document (for example links between chapters in a book). In contrast, automatic creation of content links, which connect parts of documents with similar content, is not a trivial process.

In (Furuta et al., 1989), one of the earliest work on the automatic transformation of a document into hypertext is presented. In this work, regularly and consistently structured documents, like catalog listings, traditional databases, computer programs, are transformed...
CHAPTER 2. INFORMATION RETRIEVAL AND HYPERTEXT

into hypertext. The success of this transformation depends on the degree to which there is a logical correspondence between the structured components of the documents and the generated hypertext. However, the fact that only structural links are identified makes this approach to be of less interest from the IR point of view, where content links are more important.

Rada suggests the usage of semantic nets as an intermediate form that a textbook is placed in, before its transformation into a hypertext which includes content links as well as structural links (Rada, 1992). The transformation process takes a text into first-order hypertext and second-order hypertext. The former deals with structural links, whereas the latter provides content links by connecting related terms.

Salton et al. use the information provided by the computation of the similarity between fragments of documents in order to identify content links (Salton and Buckley, 1989; Salton et al., 1993; Salton et al., 1994a). One proposed technique is to use the vector similarity to produce a network of text fragments with related content. Another approach is to use text relationship graph maps, having fragments of documents as nodes. The structure of a document can be obtained by drawing edges between the nodes of these maps whenever the pairwise similarity between the corresponding segments is above a threshold. Content links can be put automatically in case of a strong relationship between fragments of documents.

Smeaton et al. use also the similarity between document fragments. However, they selectively add links depending on an overall measure of how good the resulting hypertext is (Smeaton and Morrissey, 1995). This is accomplished by the observation of the values of various topology metrics. Compactness and stratum (see section 2.3.2) are two such metrics.

Agosti et al. suggest a conceptual architecture for information retrieval systems, structured on three levels: documents, index terms and concepts (Agosti et al., 1996). Based on this architecture, they present a methodology for the automatic construction of links between objects within each of these levels and between levels.

A recent survey about various information retrieval techniques for the automatic construction of hypertext can be found in (Agosti et al., 1997).

This thesis suggests a methodology for the automatic construction of hypertext which is tailored to the domain of newspaper archives and is based on Salton’s approach.

2.5 Conclusion

In this chapter, basic issues of information retrieval and hypertext were discussed. Information retrieval provides access to the content of documents by using direct search techniques. Content representation is performed using index terms that are derived from the documents via automatic text analysis and weighting techniques. On the other hand, hypertext provides access to the various kinds of associations that may exist between doc-
uments or their parts. Navigation and browsing help the user to discover such associations and thus to create a dynamic information path.

Hypertext information retrieval systems combine hypertext and information retrieval capabilities by using retrieval techniques that include direct search as well as navigation and browsing, in order to increase the retrieval effectiveness and provide the user with many different ways to satisfy her information need.

Thus, the incorporation of a mechanism for the automatic construction of hypertext into an IR system is quite important in order to exploit the advantage of the integration of hypertext with information retrieval.
Chapter 3

News hypertext

3.1 Overview

This chapter sets the requirements for the information retrieval on the domain of newspaper archives. Newspaper articles have special temporal characteristics and content peculiarities that should not be ignored by any IR system. The notions of stories and threads, which are substories within a story, are suggested to deal with these characteristics. A story consists of retrieved articles in response to a query. A thread connects articles with related segments by linking these segments. Stories and threads can be implemented using hypertext links, visualized by temporal layouts, and formally described using the object model.

Part of the presented work is based on (Dalamagas and Dunlop, 1997).

3.2 The newspaper paradigm

The domain of newspaper archives has been used many times in document collections\(^1\). However, traditional information retrieval systems, by presenting only a list of articles relevant to a user’s query, ignore some special characteristics concerning the newspaper domain:

- A newspaper usually presents text documents (articles) that are relatively short, and loosely related. These documents are often provided to the user with a non-linear way. For example, an overview article about political news can be in the main page, while other related articles are in subsequent pages.

- Related articles to these particular political news may be published in subsequent issues of the newspaper. Reading articles in chronological order might be of great

\(^1\)for example Wall Street Journal, Financial Times, Le Monde
importance.

- Newspaper articles tend to be *multi-narrative* and clustered by topics, as opposed to *single-narrative*, well structured documents.

For a more detailed discussion of these characteristics one can refer to (Golovchinsky and Chignell, 1997).

Taking into consideration the above characteristics, new retrieval requirements for newspaper domains may arise for the design of a retrieval system, which are presented in the following section.

### 3.2.1 Retrieval requirements for newspaper domains

- The traditional retrieval mechanism should be provided: for a query submitted by the user, the top-ranked articles are presented as relevant to the query.

- The user should be able to see the most recent or the oldest article from the list of the retrieved articles. This might be important for example in case of a trial: A user may want to read only the latest news concerning the case, having no interest about the oldest articles.

- As an extension to the previous requirement, the retrieval system should provide the user with the list of retrieved articles, ordered by the date of publication. This might be the case when the user is interested in understanding the sequence of the events. Negotiations for resolving conflicts between countries could be an example: A user might want to find out how these negotiations were evolved during a time period.

- Subtopics within the main topic which an article deals with should be identified, if present. Sometimes, for example, the cause of an air-crash accident is of increased interest to the user.

- Finally, it is quite important for a retrieval system to identify fragments within the articles that somehow summarize their content. It is common in newspaper articles to have summarizing paragraphs, especially in the beginning or the end of the articles. The identification of these fragments helps the user to find out quickly what the articles are about, without having to read them thoroughly.

The incorporation of hypertext into a retrieval system that deals with the newspaper domain, called *news hypertext*, meets the above requirements with a fair success, as this thesis suggests in the following section.
3.3 Formal aspects of news hypertext

News in printed media consist of stories which are covered in articles. Stories deal with topics that are considered to be important for the readers on the day of publication. However, a story may be a hot topic for more than one day. In that case, more than one article might be published for this story during a period of time. Note that a time gap may exist between subsequent publications of those articles. Usually, different but close aspects of a story are also examined. As a result, in a list of articles related to the story, some of them may totally or partially refer to various substories within the main story. There may also be a time gap between subsequent references to substories of a main story.

Using hypertext, articles related to a story can be linked by aggregate links (A-links). A-links are those which group together several related documents (Allan, 1996). However, for the newspaper domain, A-links also have a temporal aspect: they link pairs of related documents (articles) in a chronologically ordered chain, a story chain.

Just as articles related to a story are linked by A-links, articles that totally or partially refer to a substory can be also linked. Thread links (T-links) connect the latter ones in a chronologically ordered chain, which we call thread chain, or simply thread.

As an example, consider a news story referring to a conflict between the Scottish Secondary Teachers’ Association and the Strathclyde region that took place between May and June 1992 in Scotland. The reason of the conflict was the region’s plans to make 450 teachers redundant unless agreement could be reached on ways to achieve savings in the education budget. The story evolved through the publication of a number of related articles over a period of time. The evolution of the story started with the initial reaction of the Teachers’ Association and continued with talks between the two parts, a threat for strike, talks again, decision for strike, personal conflicts between representatives, new talks, etc.

In table 3.1, 5 articles relevant to the story are presented, ordered by date of publication. One can observe that all the segments which have been formatted with bold characters, and which belong to different relevant articles, talk about the teacher’s ballot. On the other hand, all the segments which have been formatted with italic characters refer to the talks between the two parts.

The sequence of segments that refer to the talks between the two parts is a thread of the main story of the conflict between the Scottish Secondary Teachers’ Association and the Strathclyde region. The evolution of the story is presented in figure 3.1.

Figure 3.1 also shows an overview of a story which might result from a user’s query. As such, it highlights a major difference from classic information retrieval: articles are presented as a structured subcollection and not by order of likely relevance. However, the degree of relevance can be visualized in the temporal layout, as one can see in figure 3.1. The figure 3.1 is actually the temporal layout\(^2\) of the story and it offers a simple but

\(^2\)the temporal layouts are used to describe action scenarios in multimedia applications, e.g. video and audio playback during a period of time
Table 3.1: Five articles relevant to a conflict between the Scottish Secondary Teachers’ Association and the Strathclyde region, May-Jun 1992, Scotland. One can observe the relationship between all the bold-formatted segments or between all the italic-formatted segments.
CHAPTER 3. NEWS HYPERTEXT

intuitive method to visualize the evolution of the story together with the evolution of threads, providing temporal semantics.

![Temporal layout of story evolution](image)

Figure 3.1: Temporal layout of story evolution

The object model (Rumbaugh et al., 1991) can be used to describe formally the suggested scheme, as depicted in figure 3.2. A story consists of articles and A-links which connect them. Similarly, a thread consists of articles’ segments and T-links which connect them. The notion of segments is used for the general case in which a substory is discussed only in a part of an article. A segment is considered to be a contiguous part of an article which is related to a topic that is disconnected from the adjacent text. A-links and T-links form a general object called link. Every story and thread have two basic temporal attributes based on the publication date of the first and last article contributing to the story or thread: start time and end time. Recall that during the period between the start time and the end time of a story or a thread, there may be time gaps, as one can see in the temporal layout of figure 3.1, which are not explicitly modelled in the object model. Temporal layouts as well as the suggested object model can formally set a general model to describe the attributes of news hypertext.

A conceptual model for navigating and browsing among different IR objects has been also used in (Agosti et al., 1996) (see introduction). The document level (D) of this model refers to the documents, whereas the index term level (T) refers to the index terms. The suggested concept level (C) is related to sets or classes of related index terms that are called concepts. Links may exist between objects of D and T levels or C and T levels. Also, all objects of the same level can be linked to each other (D – D, T – T, C – C). Just like a concept represents a class of index terms, a thread represents a class of articles’ segments. However, a thread is semantically enriched with the encapsulation of temporal information. In addition, thread identification can be achieved with automatic techniques,
in contrast to the manual or thesaurus-based construction of concepts’ set (Agosti et al., 1996).

The above model of threads and stories will be used in the following chapter as the framework for the development of a method for the automatic construction of news hypertext.

### 3.4 Conclusion

In this chapter, formal ways to capture the temporal aspects that characterize the newspaper domain were suggested. Aggregate links and thread links were used in order to describe the evolution of a news story. Aggregate links connect articles related to a story in a chronologically ordered chain. Thread links connect articles that totally or partially refer to a substory within the main story in some of their segments, by linking these segments.

Using such an approach, one can easily visualize the evolution of a story together with the evolution of its threads, by constructing temporal layouts. The notions of threads and stories capture the temporal aspects of newspaper domain and meet most of its requirements.
Chapter 4

NHS: News Hypertext System

4.1 Overview

This chapter discusses the framework for the implementation of the News Hypertext System (NHS), which is based on the model of threads and stories suggested in chapter 3. The major components of the NHS system are described:

- Retrieval engine: it presents to the user a list of articles relevant to a query.
- Article decomposition engine: it decomposes the articles into segments prior to any usage of the NHS system.
- Thread construction engine: it builds the threads by creating the hypertext between related segments of the retrieved articles.

Finally, a WWW based prototype of the NHS system is described.

Part of the presented work is based on (Dalamagas and Dunlop, 1997), too.

4.2 Introduction

The automatic construction of links between articles (A-links) and links between segments of articles (T-links) can be performed at index-time, prior to any usage of the system by the user, or at query-time in response to a user’s query. There are two major problems with index-time linking:

- It is potentially time-consuming and inefficient to re-examine the whole article collection for the construction of new links each time that new articles are added. In most cases, the addition of new articles does not change dramatically the structure of the hypertext.
The resulting hypertext is static, in the sense that it exists before it is used by the user and it is not adapted to the requests that she invokes each time.

As opposed to the index-time approach, this thesis suggests a methodology which is performed at query-time. The construction of links is done only for the set of retrieved articles in response to a user’s query. The suggested procedure is depicted in the *News Hypertext System (NHS)* system of figure 4.1.

This approach has the following advantages:

- The construction of hypertext is done dynamically. The resulting hypertext for a set of retrieved articles which are relevant to a query is adapted to the user’s request that is expressed through the query.

- Retrieved articles have a high probability of being relevant to the query and according to cluster hypothesis (van Rijsbergen, 1979) closely related articles tend to be
relevant to the same queries. As a result, there is also a high probability that the links connect documents that the user would consider related.

- Because of the small number of the retrieved articles, comparing with the whole article collection, clustering techniques can be performed easily and more effectively.

The suggested methodology for the automatic construction of news hypertext is summarized as follows:

- Decompose all the articles of the collection into segments, prior to any usage of the NHIR system.
- For all the retrieved articles in response to a user query, apply clustering to their segments.
- Link the retrieved articles in a chronologically ordered chain in order to create a main story which is relevant to the query.
- Link the segments that belong to the same cluster in a chronologically ordered chain in order to create a thread, which refer to a substory within the main story.

The components of the NHS system, the basic retrieval engine, the article segmentation engine and the thread construction engine, which are used for the implementation of the above methodology, are discussed in more details in the following sections.

### 4.3 Retrieval engine

The SMART retrieval system (Salton and McGill, 1983; Buckley, 1985) was selected as the retrieval engine in which the suggested framework was implemented and tested. Although the SMART system is an academic research software and it is not optimized for any particular usage, it has been designed with great flexibility.

The SMART system follows a multi-level architecture that implements the vector space retrieval strategy (see chapter 2). It is composed of four levels of programs and procedures, which make it a fully automatic document retrieval system:

1. *User request level*
   
   A user submits her requests to the user request level. Typical requests include: indexing a collection, indexing and adding new documents in an already indexed collection, retrieving documents in response to a query, performing experimental retrieval and evaluation, etc.
2. **Task implementation level**
   According to the user’s request, the system decides which tasks need to be performed. The task implementation level carries out these tasks using various modules. The *indexing module* transforms a document or a query into a set of index terms and concepts. The *retrieval module* matches an indexed query to all the indexed documents of the collection and retrieves relevant documents in response to that query. The *display module* is responsible for presenting the retrieved documents to the user. The *feedback module* performs relevance feedback. Finally, the *evaluation module* gives results for the retrieval effectiveness, based on precision-recall values.

3. **Object access level**
   During the performance of the tasks in the task implementation level, the access to data is through *file objects*. Each object has been implemented as a set of data structures and procedures that deal with that structures. The object access level contains the procedures for accessing file objects and reading the *specification files*. The specification files contain parameter information that many SMART procedures need.

4. **Database access level**
   File objects have been implemented in database access level using well-known database access methods, like hashing, sorting, etc. File objects include *dictionaries, inverted files*, the well-known in vector space retrieval document (or query) *vector*, etc.

The document collection which was used with the SMART retrieval system, for the NHS system, consists of 23209 articles (100MB) from “The Herald” newspaper (Jan 1992 - Jun 1992). “The Herald” is a Scottish broadsheet that covers a wide range of news (economy, politics, local news, sports, social and culture issues etc.) and is not biased towards any particular subjects or issues. This gives access to a large local user base. The collection was indexed\(^1\) using the normalized raw \(tf \cdot idf\) weighting formula (see chapter 2).

Having the SMART as the underline retrieval engine, the NHS system retrieves a set of articles that are considered relevant to a user’s query. In addition to the traditional way of presenting a ranked list of these articles to the user, the NHS presents them also in a chronologically ordered chain, in order to create a main story which is relevant to the query. This chain is the source for constructing the links between the articles (*A*-links).

\(^{1}\)indexing speed: 9min for 100MB, using a publicly available SPARCstation-20, with 2CPU’s (75MHz each) and 256MB RAM (load avg 2.5)
4.4 Article decomposition

Creating links between the articles (A-links) that have been retrieved as relevant to a query is straightforward. The articles are connected via links in a chronologically ordered chain in order to form a story.

As opposed to the easy construction of A-links, creating thread links (T-links) is a process which initially needs text decomposition so that segments of articles are identified. Recall that a segment is a contiguous part of an article which is related to a topic that is disconnected from the adjacent text. This topic may refer to a substory within the main story. The suggested procedure for text decomposition has been successfully explored by Salton et al. and is described in the following section (Salton et al., 1994b; Salton et al., 1995).

4.4.1 Segment detection based on paragraph relationship maps

A paragraph relationship map of a document has the form of a graph. Its vertices correspond to the paragraphs of the document whereas its edges refer to links between the paragraphs. Similarity measures between pairs of paragraphs are usually put as labels in the edges. Such a map is presented in figure 4.2a. In this figure, for example, paragraph p1 is related closely to paragraph p2 because their similarity measure is 0.8. In contrast, a similarity measure of 0.3 between paragraphs p2 and p3 shows that these paragraphs are not related.

The text decomposition procedure is described as follows:

1. Construct the paragraph relationship map of the document that needs to be decomposed into segments (see figure 4.2a).
2. Drop the links of the map that correspond to similarity values which are below a predetermined threshold (see figure 4.2b).
3. Drop long-distance links: those spanning more than a predetermined number of adjacent paragraphs (see figure 4.2c).

At the end of this procedure, a break down into separate sets of connected paragraphs is expected, which results in segment formation. For example, in figure 4.2c, 2 segments have been detected. The first segment consists of p1, p2 and p3 paragraphs and the second one of p4, p5 and p6 paragraphs.

The above procedure actually identifies connected components\(^2\) of the paragraph relationship map (graph). The main characteristic of these connected components is the lack of long-distance links.

---

\(^2\) A connected component of a graph is a set of vertices such that each vertex is connected to at least one other vertex and the set is maximal with respect to that property.
4.4.2 Implementation issues

The above methodology for text decomposition has been implemented by performing the following procedures for each one of the articles in the collection, prior to any usage of the NHS system:

1. The paragraphs of an article are identified and indexed as separate documents, using the SMART system. Because of the fact that the articles consist of short paragraphs with fairly homogeneous length, unnormalized $tf \cdot idf$ weighting scheme is used, so that the text similarity depends on the number rather than the proportion of matching terms (Salton et al., 1994a).

2. Similarity values are computed for each pair of paragraphs, using the SMART system. Low similarity values are not taken into account. Similarity values are considered to be low if they are below a threshold $thr$, which is set to be half of the maximum similarity value that is detected between two paragraphs. Although the approach is a heuristic one, it seems to work effectively because too high similarity values, that are well distinguished from the rest and thus could bias the estimated $thr$ to undesirable high values, have not been observed.

3. Similarity values between long-distant paragraphs are not taken into account, too. A distance $D$ between two paragraphs is an indication of how far away is a paragraph from another one within an article. If an article consists of $m$ paragraphs $p_1, p_2, \ldots, p_m$, with $p_1$ ($p_m$) denoting the first (last) paragraph of the article, the distance $D$ between two paragraphs $p_i, p_j$ is defined as:

$$D(p_i, p_j) = |i - j|, \quad i, j = 1, \ldots, m$$  \hspace{1cm} (4.1)

For example, two adjacent paragraphs have distance equal to 1. Two paragraphs are considered to be long-distant if their distance is greater or equal to 3. Again the
approach is a heuristic one and is based on the observation that acceptable similarity
values between two paragraphs that have more than 1 paragraph to exist between
them within the article are not actually a result of content similarity.

The remaining paragraph-to-paragraph similarity values correspond to the paragraph re-
relationship map of the specific article. These values are examined so that chains of adja-
cent linked paragraphs are identified within the article. A set of $k$ adjacent paragraphs
$p_1, p_2, \ldots, p_k$ within the article, with $p_1$ ($p_k$) denoting the first (last) paragraph of the set,
is considered to be a chain if:

\[
(p_i, p_{i+1}) \in SV, \quad \forall i, \quad i = 1, \ldots, (k - 1)
\]

(4.2)

where $SV$ is the set of all the pairs of paragraphs that form the paragraph relationship
map. These chains are a first indication of possible segment formation. One example of
such a chain is the set of $p_4 - p_5 - p_6$ paragraphs in the map of figure 4.2c.

These chains are extended in order to include distant paragraphs ($D < 3$). The two
primitive schemes that refer to distant linked paragraphs are presented in figure 4.3. As
one can see in figure 4.3, the primitive schemes are always transformed into chains of
adjacent paragraphs. The fact that two paragraphs are related to the same one, is an
indication that all three paragraphs are somehow related. Thus, the initial chains may be
extended if one of the two primitive scheme is found at the beginning or the end of a chain.
This is depicted clearly in figure 4.4. The constructed chains of paragraphs correspond to
segments within the article.

Following the above three steps, all the articles of a collection can be decomposed into
segments at index-time, prior to any usage of the NHS system. After this process, $n_i$
segments exist for each article $a_i$:

\[
a_i \rightarrow (S_{i1}, S_{i2}, \ldots, S_{in_i})
\]

(4.3)
Figure 4.4: Extending chains of adjacent paragraphs: Chain CH1 initially consists of paragraphs $p_1$, $p_2$, $p_3$, $p_4$. Because a primitive scheme of distant linked paragraphs is detected (see $p_3$, $p_4$, $p_5$), chain CH1 is extended to chain CH in order to include also $p_5$. Similarly, chain CH2 is extended to chain CH in order to include also $p_4$, $p_5$, because of the primitive scheme $p_3$, $p_4$, $p_5$.

After the segments of each article have been identified, they are indexed as separated documents using the SMART system. Because the length of the segments may vary, normalized augment $t_f \cdot idf$ weighting scheme is used. The indexes are needed for the computation of similarity values, as it will be explained in the next section.

### 4.5 Thread construction

Having the set of all segments $S$ of all articles, clustering can be performed for the segments of the retrieved articles relevant to the query. Each formed cluster is used to create a thread. Links are put to connect the components (segments) of each cluster in a chronologically ordered chain. Following this way, threads within the story are constructed.

The automatically created hypertext has the form that is depicted in the example of figure 4.5. Despite the fact that the articles in figure 4.5 are not presented by order of likely relevance, highly ranked articles are usually grouped together in a thread. This is the case when the user inserts query terms that are relevant to a topic which a thread refers to. Performing clustering for articles’ segments can also deal with problems of non-sequential topic treatment. This can happen when the subtopics within an article are not well isolated from each other. In that case, the described text decomposition process extracts segments that may be related to each other, although they are separated in the article. Clustering
4.5.1 Clustering methods

Clustering methods are usually divided into two broad categories. Non-hierarchical methods divide a data set into a number of clusters. Hierarchical methods produce nested sets of data (hierarchies), in which pairs of elements or clusters are successively linked until every element in the data set becomes connected. Non-hierarchical methods have low computational requirements, $O(kn)$ (if for example $n$ documents need to be grouped into $k$ clusters), but certain parameters like the number of formed clusters must be known a priori. Hierarchical methods are computationally expensive, with time requirements of $O(n^2)$ (if $n$ documents need to be clustered), because all pairs of similarities between the documents must be considered\(^3\). However, hierarchical methods have been used extensively as a means of increasing the effectiveness and efficiency of retrieval (Jardine and van Rijsbergen, 1971; Voorhees, 1985; Hearst and Pedersen, 1996). For a wide ranging overview of clustering methods one can refer to (Willet, 1988; Rasmussen, 1992).

\(^3\)Willet (Willet, 1980) suggests an inverted file algorithm to limit the amount of computation required to calculate a similarity matrix

Figure 4.5: News hypertext

will also group together these segments.

As an alternative to clustering, Allan suggests to use the similarity matrix between document segments and put links between segments that are related with high similarity value (Allan, 1995). This approach does not take into consideration the temporal aspects of news hypertext and the notion of story evolution. The clustering approach which is suggested in this thesis is more appropriate in case of news articles, because a cluster of articles’ segments can be used to represent a substory within a main story that a set of articles is related to. Clusters, transformed into chronologically ordered chains, provide a better way to deal with the notion of story evolution.
Single link, complete link and group average link compose the group of hierarchical clustering methods. All of these different methods can be described by the following algorithm:

1. Each element of the data set to be clustered is considered to be a single cluster.

2. The clusters with the maximum similarity are merged and the similarity between the remaining clusters and the new, merged one is recalculated.

3. While there are more than one clusters, go again to step 2.

In single link (complete link), the similarity between two non-single clusters is defined as the maximum (minimum) of the similarities between all pairs of elements so that one element is in the one cluster and the other element is in the other cluster. In group average link, the similarity between two non-single clusters is defined as the mean of the similarities between all pairs of elements so that one element is in the one cluster and the other element is in the other cluster.

Single link has been shown to be theoretical sound, under a certain number of reasonable conditions (Jardine and Sibson, 1971; van Rijsbergen, 1979).

4.5.2 Implementation issues

In any case, the choice of an appropriate clustering algorithm is of great importance for the quality of the created $T$-links and the performance of the system.

An interesting approach to the problem of on-the-fly clustering can be found in the Scatter/Gather browsing method suggested in (Cutting et al., 1992). Following this approach, group average agglomerative clustering is used for a small random sample of documents in order to find cluster centroids. Since $k$ centroids have been found, each document is assigned to one of these centroids. Refinement procedures may follow.

For the NHS system, the single link clustering method is used. Two are the main reasons for making this selection:

- Initial experiments with single-link method showed that it does not impose great performance overhead on the NHS system. More details follow in subsequent paragraphs.

- The single link merges the most similar clusters into a new one with each element being in the same cluster with its nearest neighbour. Applying this to the segment clustering paradigm, each segment is in the same cluster with its nearest neighbour, thus with the segment which is the most similar to it. This is quite close to the notion of thread.
The single link clustering which the NHS systems uses is based on Prim’s algorithm (Cormen et al., 1990) for computing the maximum spanning tree (MST) of a graph. Given a graph $G$ with a set of weighted edges $E$ and a set of vertices $V$, a MST is an acyclic subset $T \subseteq E$ that links all the vertices and whose total weight $W(T)$ (the sum of the weights for the edges in $T$) is maximized. It has been shown (Gower and Ross, 1969) that a MST contains all the information needed in order to perform single link clustering.

The $n$ segments of the retrieved articles form a graph $G$ with $n$ vertices $v \in V$ and $n(n - 1)/2$ weighted edges $e \in E$. The weight of an edge corresponds to the similarity value between the vertices (segments) that this edge connects. The Prim’s algorithm for MST construction proceeds as follows:

1. $A$ is the set, initially $\emptyset$, that contains the edges assigned to the MST, while $B$ is the set, initially $E$, that contains those not assigned to the MST.
2. For a given vertex $v$, assign to $A$ the maximum weighted edge $(v, u) \in B$.
3. Continue to add to $A$ the maximum weighted edge $(v_i, u_j) \in B$, which connects to at least one edge from $A$ without forming a closed loop amongst the edges already in $A$.
4. The algorithm stops when there are $n - 1$ edges in $A$.

Given the MST, the single link clusters for a weight (similarity) level $l_1$ can be identified by deleting all the edges from the MST with weight $w < l_1$. The connected components of the remaining graph are the single link clusters. The process is depicted in figure 4.6.

Three are now the problems that arise in partitioning the set of segments into clusters:

1. computation of similarity values for each pair of segment
2. connected component identification
3. selection of the most appropriate similarity (weight) level for clustering

Calculating similarity values

Similarity values between all pairs of segments are computed using the SMART system, which has been modified in order to perform this process automatically after the retrieval of the articles, based on the indexes of the segment collection (see previous section).

Determining connected components of a graph

Disjoint-set data structure operations are used for determining connected components of a graph (Cormen et al., 1990). A disjoint-set data structure is represented by a collection
Figure 4.6: Maximum Spanning Tree (MST) detection and single link clustering: (1)-(7): the execution of Prim’s algorithm starting from vertex A (the dashed lines are the edges that are added to the MST at each step of the algorithm), (8): MST, (9): clusters at level 0.6

$S = (S_1, S_2, \ldots, S_k)$ of disjoint sets. Given that $O$ is an object, the following operations are supported for the collection $S$:

- $MAKE-SET(O)$: creates a new set $S_O$ whose only member is $O$.
- $UNION(O_1, O_2)$: unites the two sets $S_{O_1}, S_{O_2}$ that contain $O_1$ and $O_2$, respectively.
- $FIND-SET(O)$: returns the set that contains $O$.

Using the above operations, the connected components of a graph can be identified as follows:

1. for each vertex $v \in V$ $MAKE-SET(v)$
2. for each edge \((v, u) \in E\): if \(\text{FIND-SET}(v) \neq \text{FIND-SET}(u)\) then \(\text{UNION}(v, u)\)

Selecting the clustering level

The goal of applying single link to the set of segments of the retrieved articles is to partition this set into a number of clusters. This number is not known a priori. Thus, a stopping rule must be applied in order to determine the most appropriate clustering level for the single link hierarchies. Milligan et al. present 30 such rules (Milligan and Cooper, 1985). In the NHS system, a simple approach is followed. The single link clustering process is performed for all similarity levels between \(0.7\) (upper bound) and \(0.3\) (lower bound), with step \(0.02\). The lower bound was selected by observation, so that links between segments with low similarity values are not taken into account. The upper bound was set after observing that similarity values between segments which are higher than \(0.7\) are quite rare.

<table>
<thead>
<tr>
<th>CLUSTERING LEVEL</th>
<th>NUMBER OF EDGES</th>
<th>NUMBER OF CLUSTERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0.68</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>0.66</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>0.64</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>0.62</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>0.6</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>0.58</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>0.56</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>0.54</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>0.52</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>0.5</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>0.48</td>
<td>18</td>
<td>10</td>
</tr>
<tr>
<td>0.46</td>
<td>22</td>
<td>10</td>
</tr>
<tr>
<td>0.44</td>
<td>24</td>
<td>10</td>
</tr>
<tr>
<td>0.42</td>
<td>27</td>
<td>9</td>
</tr>
<tr>
<td>0.4</td>
<td>32</td>
<td>8</td>
</tr>
<tr>
<td>0.38</td>
<td>39</td>
<td>10</td>
</tr>
<tr>
<td>0.36</td>
<td>45</td>
<td>9</td>
</tr>
<tr>
<td>0.34</td>
<td>50</td>
<td>11</td>
</tr>
<tr>
<td>0.32</td>
<td>58</td>
<td>13</td>
</tr>
<tr>
<td>0.3</td>
<td>65</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 4.1: Refinements of a single link clustering process

The \(k\) clustering levels \(l_1, l_2, \ldots, l_k\) \((l_1 > l_2 \ldots > l_k)\) are considered to be good to start with, if for all values \(l_i, i = 1, 2, \ldots, k\), the clusters remain the same, whereas for values
slightly higher than $l_1$ clusters are broken up. The NHS uses the last level just before the breaking of the clusters as a candidate, in order for the clusters to be as inclusive as possible. For example, in table 4.1 one can observe that the levels 0.6, 0.58, 0.56 are good ones to start with, because the clusters remain the same for two refinement steps (0.58, 0.56). However, NHS will use the last level before the changing of the clusters, thus level 0.56. Similarly, the levels 0.48, 0.46, 0.44 give as another possible final selection the level 0.44 (2 refinement steps: 0.46, 0.44). The level 0.62 is also selected as the last level before the changing of clusters from 4 to 6 (4 refinement steps: 0.68, 0.66, 0.64, 0.62). Because the level 0.62 comes from more refinement steps than the other two levels, it will be used finally as the clustering level.

The performance of single link clustering does not impose great overhead on the system. Clustering of 200 segments takes less than 2 min, including the computation of similarities, using a publicly available SPARCstation-20 with 2CPU’s (75MHz each) and 256MB RAM (load avg 2.5). 200 segments correspond to more than 100 articles.

4.6 NHS prototype

A working prototype of the NHS system has been implemented\textsuperscript{4}. Its structure extents the three-level architecture that was presented in the previous paragraphs by adding a WWW interface\textsuperscript{5} on the top of the levels. Most of the users are nowadays familiar with WWW browsers and this makes the NHS system easy and simple to use. Diagrams of the components of the NHS prototype system are depicted in figure 4.7.

As one can observe in that figure, the thread construction is the result of segment clustering. After a set of articles have been retrieved in response to a user’s query (see figure 4.7: RETRIEVAL ENGINE), similarity values are calculated for each pair of their segments. The segments for each article are easily found by using the inverted file ARTICLE-SEGMENTS, which has been created during the decomposition process, prior to any usage of the NHS system (see section 4.4). The indices of the segment collection provide the source for a modified version of the SMART system to calculate the pairwise similarities (see figure 4.7: THREAD CONSTRUCTION ENGINE). It must be noticed that only the indices of the segment collection are needed and not its whole text. The calculated similarity values are stored in a file, which is used by the clustering process in order to determine groups (clusters) of related segments. For each cluster of related segments, their corresponding articles are found using the SEGMENT-ARTICLE inverted file. The linking information, that is the location of the link in an article and its direction, are stored in the ARTICLE-LINKS inverted file. Linking information concerning each thread separately is stored in the THREADS file. The aforesaid two files are the source for creating the hypertext during user’s navigation using the WWW interface (see figure 4.7: WWW connection).

\textsuperscript{4}PERL script language, C language and CGI scripts are the software tools that were used
\textsuperscript{5}screen snaps of the WWW interface of the NHS system can be found in the appendix
4.7 Conclusion

In this chapter, the NHS system which automatically creates hypertext links for news hypertext was presented. The NHS system retrieves news articles relevant to a query and presents a story with threads as a result. A story is created by linking the articles in a chronologically ordered chain. Threads are constructed by linking related segments of articles in the hope of capturing different substories. Threads also form chronologically ordered chains.

Related segments are identified by applying clustering techniques to all the segments of the retrieved articles. The clustering process, which is the core of the hypertext construction, is done on-the-fly and only for the set of the retrieved articles, without imposing great overhead on the system. This approach has the advantage of constructing hypertext dynamically by adapting it to the user’s request which is expressed through her query.

A working prototype NHS system was also presented. Because of its WWW based interface, it provides a simple and easy way for the users to test its effectiveness.
Chapter 5

Evaluation

5.1 Overview

In this chapter methods to evaluate the effectiveness of the NHS system are presented. The evaluation procedure was performed in three parts. Firstly, the quality of the decomposition process was tested, by carrying out users experiments. Thread evaluation based on hypertext metrics as well as user tests was also performed. Finally, the overall effectiveness of the NHS system was examined, by designing a user experiment in which the NHS system was tested in comparison with a normal IR system.

5.2 Introduction

Information retrieval evaluation techniques, like precision and recall graphs, cannot be directly used for the evaluation of the NHS system. The presented methodology for the automatic construction of news hypertext should be evaluated from three different aspects:

- **Hypertext structure analysis**
  Various metrics have been developed for structural analysis of hypertext. Compactness and stratum (see chapter 2) can be used to measure the overall topology of a hypertext and therefore to provide some evidence about its quality (Smeaton and Morrissey, 1995). For example, high-compact hypertext, where lots of links exist, usually leads to the well-known problem of disorientation (Conklin, 1987). In case of news articles, users usually look for flat hypertext structure. Chains of linked articles relevant to a topic are helpful, as long as they are not extremely long. Inter-chain links, that is links between different chains of articles, generally lead to disorientation, although sometimes are useful to discover related topics.

- **Hypertext quality**
  Clustering determines threads of articles that refer to substories within the main
story and therefore affect the quality of the resulting hypertext. Cluster validity
techniques can help in order to identify whether the set of the texts which a thread
consists of exhibits a clustering tendency or not. An example is the usage of random
document graphs (Ling and Killough, 1976). In this approach, a set of \( n \) documents
is represented with a graph of \( n \) vertices and \( v \) edges. Each vertex corresponds to a
document. An edge connects a pair of documents (vertices) only if their similarity
exceeds a threshold. The likelihood of this structure occurring randomly can be
estimated by comparing it to random structures of graphs with \( n \) vertices and \( v \)
edges. For an overview in these techniques one can refer to (Willet, 1988).

Hypertext quality is also affected by the decomposition process that is used in order
to identify segments within an article. The effectiveness of this process should also be evaluated.

- **System usability**
  The effectiveness of the NHS system needs also to be tested from the user’s point
  of view. The time that a user spends in order to fulfil her information needs is a
critical measure for the effectiveness of the NHS system. Therefore, a comparison
between a classical IR system and the NHS system can be made on the basis of the
required time for the fulfilment of the users’ information needs. The user should
perform a certain set of tasks which involve expressing her information need with
the form of a query, navigating within the threads and deciding the relevance of the
article. The system will be characterized as an effective one if the user fulfils her
information needs after a short-time navigation. Discussion tools based on human-
computer interaction techniques, like user questioning and think-aloud protocols,
may also be used. For a discussion of the issues in interactive evaluation, including
hypertext information retrieval, the reader is referred to (Dunlop, 1996).

Based on the above remarks, the evaluation procedure was performed in three parts: segment evaluation, thread evaluation and overall evaluation. The following sections discuss each part in detail.

### 5.3 Segment evaluation

Segment evaluation was based on user tests. 10 segmented articles with their segments marked were given to 10 users. The users were asked to evaluate the quality of the decomposition process by answering questions of the following form:

- Do the sets of paragraphs that are marked form a real segment?
- Are there more segments that were not found?
- Do the sets of paragraphs that are supposed to form a segment contain more (or less) paragraphs that are really needed?
CHAPTER 5. EVALUATION

Table 5.1: Segment evaluation results from user experiments

- Are you satisfied with the results?

As one can see in table 5.1,

1. the segments were in general well identified
2. failure to detect all segments occurred only in a few cases
3. the boundaries of the segments were sometimes not well defined
4. the users were generally satisfied with the overall results

The suggested method for text decomposition into segments seems to have fair success, despite the fact that sometimes times the boundaries of the segments are not correctly identified. This is the case when more paragraphs (usually only 1) are included in the segment although they are not needed or when some paragraphs (usually only 1) are not contained in the segment although they are needed. This happens because the NHS decomposition engine does not verify the decision for segment detection by looking also at text units smaller than paragraphs (e.g. sentences). See for example the decomposed article about wines that is presented in table 5.2.

5.4 Thread evaluation

5.4.1 Hypertext metrics

Five queries that correspond to five different long news stories were manually constructed. The news stories were picked out randomly with the following methodology:

- Pick up an article.
Almost every producer is desperate to sell, and poor wine is rare. Most of the earth’s surface is covered by water, but you might be forgiven for imagining it is really wine when there is so much of it about. And the worldwide recession (well you do not really believe that nonsense that it is here only, and all the work of Lawson, Major and Lamont, did you?) means that it should be cheap. Almost every producer is desperate to sell, and poor wine is rare, though importers are sad to find that Czech wine is not better than it is and that Portuguese products (other than port) show some signs of decline. But decent wine around 3 is fairly common, with Marks and Spencer’s Côtes de Gasgoine and Asda’s Domaine de Grangeneuve providing a handy white and red respectively for inexpensive casual tippling.

It is still necessary to pay a bit more for something better, and single-vineyard wines have generally managed to hold their prices, though I am told that a testing time in the UK auction houses may arrive this Easter. Since these deal overwhelmingly in top and near top-quality stuff, there may be buying opportunities later.

Sometimes a single vineyard wine can be relatively cheap. One such example is Chateau Loudenne blanc which is available from Peter Dominic’s (there is one in Bearsden) at just over 5. A gap has opened up between this wine and its red sibling, which is now a good bit dearer, though produced on the same location, if, mainly, from vines on the better slopes.

The vineyard is famous because of its history – it has been English-owned for over a hundred years – and partly because its present owners have made it a showpiece and run an interesting "academy" there which offers an inexpensive summer school holiday for wine lovers. The owners, ultimately Grand Metropolitan, use it to produce wines to the highest standards. That they are not better than they are is a matter of geology, of location rather than anything else.

The red is always an approachable wine because it has a lot of merlot, often 50%, in its mix to offset the austerity imposed by the cabernet sauvignon. Its price reflects the fact that the growing areas are all within the limits entitled to the appellation Haut Medoc. The white grapes, sauvignon blanc and semillon, enjoy no such advantage; they are not, generally, on the best locations and these are often not within the aforesaid limits – their wine is a mere Bordeaux sec, a designation which covers a multitude of sins.

The fact remains that Chateau Loudenne itself is firmly in the Medoc and its white wine is one of only three that I know within that famous region (there may be others now, but hardly many.) The two others are rare enough; that made at Chateau Margaux is dear; that at Chateau Talbot (called Cailloux blanc de Chateau Talbot) turns up occasionally. It is somewhat better and with a fuller flavour than Loudenne, but when last encountered in Glasgow (at the now-defunct Agnew’s) was a good bit dearer. In short, Loudenne blanc was one of only three white Medocs, and the most affordable – which is why I’ve gone on about it before. It is also available.

Table 5.2: SECTION 1: introduction, SECTION 2: vineyard wines (observe that the paragraph that follows the end of the second segment should have been included in that segment since it talks about the history of vineyard wines), SECTION 3: other varieties produced
CHAPTER 5. EVALUATION

Table 5.3: News stories

- Use the article as a query and retrieve relevant documents.
- If there are many retrieved articles with high score, which is an indication of the presence of a news story, use this article as a possible source of keywords in order to construct a query.

The selected queries cover a wide range of news (social, economic, international etc), as one can see in table 5.3. The number of articles retrieved from the system is based on a cut-off that was set in order to discard articles with low score (below 0.1). The NHS used the above queries in order to detect threads within the news stories and create hypertext.

Generally, the hypertext that the NHS creates is expected to have the form of a cyclic graph with gaps, as one can see in figure 5.1. This reflects the notion of threads, which are distinguishable and well-separated sets of related segments. This specific structure results in low values of compactness and stratum, as one can see in figure 2.3 of chapter 2.

The calculated values of compactness and stratum of the resulting hypertext for the five aforementioned queries are presented in table 5.4. The results confirm the expectation for the hypertext structure, giving an average of 0.14 for compactness and 0.15 for stratum.
### Table 5.4: Compactness and stratum for the resulting hypertext of the five stories

<table>
<thead>
<tr>
<th>NUMBER</th>
<th>COMPACTNESS</th>
<th>STRATUM</th>
<th>LINKS/NODES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.13</td>
<td>0.14</td>
<td>10/14</td>
</tr>
<tr>
<td>2</td>
<td>0.12</td>
<td>0.025</td>
<td>13/19</td>
</tr>
<tr>
<td>3</td>
<td>0.12</td>
<td>0.11</td>
<td>5/8</td>
</tr>
<tr>
<td>4</td>
<td>0.17</td>
<td>0.25</td>
<td>2/4</td>
</tr>
<tr>
<td>5</td>
<td>0.16</td>
<td>0.21</td>
<td>9/12</td>
</tr>
<tr>
<td>AVG</td>
<td>0.14</td>
<td>0.15</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 5.4: Compactness and stratum for the resulting hypertext of the five stories

### Table 5.5: Average marks from 10 users (U1, ..., U10) of the quality of threads for 5 queries (Q1, ..., Q5). GRADE SCALE: 1, 2, 3, 4

<table>
<thead>
<tr>
<th>U1</th>
<th>U2</th>
<th>U3</th>
<th>U4</th>
<th>U5</th>
<th>U6</th>
<th>U7</th>
<th>U8</th>
<th>U9</th>
<th>U10</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4</td>
<td>2.5</td>
<td>3.0</td>
<td>2.8</td>
<td>3.2</td>
<td>2.8</td>
<td>3.5</td>
<td>3.5</td>
<td>3.3</td>
<td>3.3</td>
</tr>
<tr>
<td>Q1</td>
<td>Q1</td>
<td>Q2</td>
<td>Q2</td>
<td>Q3</td>
<td>Q3</td>
<td>Q4</td>
<td>Q4</td>
<td>Q5</td>
<td>Q5</td>
</tr>
</tbody>
</table>

Table 5.5: Average marks from 10 users (U1, ..., U10) of the quality of threads for 5 queries (Q1, ..., Q5). GRADE SCALE: 1, 2, 3, 4

### 5.4.2 User judgements

The quality of links were evaluated using judgements from 10 users. For each query of those that were presented in the previous section, 2 different users marked the linking quality using a four-grade scale (1: very bad, 2: bad, 3: good, 4: very good). Table 5.5 presents the evaluation results. As one can observe, the linking is generally characterized as “good”, with marks quite close to 3.

Apart from the marking, interesting results were obtained by applying the think-aloud protocol during the evaluation, asking the users to make comments while they were marking. The following conclusions were deduced:

1. Users were satisfied even if there were a few false links. From their point of view, the short length of the linked segments (comparing to the length of a full article) and the small number of articles that a thread usually contains, helped them to easily detect the false links and not take them into account while browsing. An example is depicted in table 5.6.

2. Many threads consists of quite similar segments. This is the case when summarizing paragraphs which are repeated in many articles, usually at the beginning or the end of the article, have been identified as segments. One such thread is presented in table 5.7.

3. Non relevant threads, that is threads that link segments which belong to non relevant articles, are rare (details in the subsequent section).

Examples of detected threads are depicted in table 5.8.
**CHAPTER 5. EVALUATION**

<table>
<thead>
<tr>
<th>ARTICLE NO: 17550</th>
<th>TITLE: Strike threat as teachers face redundancies</th>
<th>DATE: 21/05/1992</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A ONE-day teachers’ strike in Strathclyde fol-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>lowed by a programme of industrial action was</td>
<td></td>
</tr>
<tr>
<td></td>
<td>threatened by the Educational Institute of Scot-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>land yesterday.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The move was in response to the Labour adminis-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tration’s proposal to make 450 full-time teach-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ers redundant.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The national executive of the EIS, which rep-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>resents 19,000 of the 23,000 teachers in the</td>
<td></td>
</tr>
<tr>
<td></td>
<td>region, agreed yesterday to ballot its mem-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>bers.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The region has already started discussions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>with EIS officials on the criteria for selection of those who would be made redundant.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mr Jim Martin, EIS general secretary, said that although they would continue to discuss the proposed redundancies with the regional education authority, they had decided to combat the prospect of job losses with a strike ballot.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“This has not happened in Scotland before,” he said. “It is the first time any local authority has gone so far down the line to make full-time permanent staff redundant.”</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Councillors and representatives of the education department will meet EIS officials today to try to reach agreement...</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ARTICLE NO: 20570</th>
<th>TITLE: Teachers up before head for their jobs</th>
<th>DATE: 28/05/1992</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>... Schools would be required to revise the staff pay complement. In effect, head teachers in primary and secondary schools are to be given revised staffing complements and nominate teachers surplus to requirements.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Councillor Ian Davidson, chairman of the educa-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tion committee, said it was part of an on-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>going exercise.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The teachers had said that any redundancies should be broadly on a basis of last in, first out. That might well save the region money, but was not considered the appropriate way to proceed, he said.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>It is expected that the head teachers will be giving their lists of redundancies to the region by June 10, which is the day after a threatened one-day stoppage by members of the Educational Institute of Scotland. The union is opposed to any form of compulsory redundancy among teachers.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Talks aimed at resolving the crisis failed to achieve agreement yesterday and are to be resumed on Monday.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mr Drew Morrise, an official of the EIS, said the union was still discussing technical details with the council’s officials.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ARTICLE NO: 17311</th>
<th>TITLE: EIS in talks with region over cuts</th>
<th>DATE: 22/05/1992</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Councillors and representatives of the Educa-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tional Institute of Scotland met yesterday to identify areas in which the savings could be made. The authority believes they could come largely from an extension of the current agreement on absence cover.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>However, other areas have been suggested, in-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cluding a freeze on recruitment, voluntary se-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>verence, and reduced spending on the delegated management of resources scheme and the quality assurance unit.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>They will all be considered in greater detail at next week’s meeting. Strathclyde has warned that about 400 teachers will be made redundant unless agreement can be reached on the savings. The leader of Strathclyde region, Councillor Charles Gray, described yesterday’s talks as reasonably amicable.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ARTICLE NO: 19743</th>
<th>TITLE: Teachers vote for strike</th>
<th>DATE: 04/06/1992</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TEACHERS have voted by a margin of two to one to take industrial action against the threat of up to 450 compulsory redundancies in Strathclyde region.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A poll of teachers belonging to the Educa-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tional Institute of Scotland showed 67% sup-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>porting strike action and 33% against.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The general secretary of the EIS, Mr Jim Ma-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rtin, said yesterday: “I very much welcome the support for action by members in Strathclyde”.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>More talks were held yesterday between of-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ficers of Strathclyde education department and EIS officials to identify cuts of more than 5m in the education budget. They are to resume on Monday.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Several cost-cutting options have been ident-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ified and if these are palatable to both sides, the dispute could be settled next week.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The ballot result, the progress of the talks, and the strike option will be discussed in the EIS executive today.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ARTICLE NO: 21248</th>
<th>TITLE: Teachers vote on strike over threat to sack 450</th>
<th>DATE: 25/05/1992</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>THE Scottish Secondary Teachers’ Association is to ballot its 2700 Strathclyde members on possible strike action in protest at the region’s plans to make 450 teachers redundant unless agreement can be reached on ways to achieve savings of more than 5m in the education budget.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The teachers will be asked to vote on a one-day strike on June 9, the same day being considered for action by the Educational Institute of Scotland, which represents 19,000 of the 23,000 teachers in the region.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The result of the EIS ballot is expected on Wednesday. If the vote is in favour a one-day strike will be followed by industrial action throughout next month, which could be carried into August and September.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mr Alex Stanley, SSTA secretary, said yesterday: “Strathclyde region’s planned cuts are totally unnecessary. The region has said one of its reasons is because there has been a reduction in demand for education.”</td>
<td></td>
</tr>
<tr>
<td></td>
<td>That is an absolute nonsense. School rolls are actually rising, they are not declining.”</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mr Stanley said the executive of the SSTA would meet next week to plan further action that should prove necessary.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ARTICLE NO: 19534</th>
<th>TITLE: Prospects of an end to teachers’ row</th>
<th>DATE: 05/06/1992</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PROSPECTS of a settlement in the dispute between Strathclyde region and the teachers’ unions over the threat of 450 compulsory redundancies improved yesterday.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The national executive of the main teachers’ union, the Educational Institute of Scotland, unanimously agreed to suspend strike action to allow negotiations aimed at reaching an agreement on avoiding teacher redundancies to take place on Monday.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The near two-thirds of EIS members in the region who took part in a ballot on industrial action voted by a margin of two-to-one in favour of striking.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Left-wing delegates attending the union’s annual conference in Perth yesterday expected the executive to identify a day on which a strike would take place next week.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>However, the executive decided to hold back on industrial action and agreed that national officials of the union should take part for the first time in talks with the authority.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The decision angered those Strathclyde officials and delegates who wanted to pull out of the talks and go on strike.</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.6: Six articles relevant to a conflict between the Scottish Secondary Teachers’ Association and the Strathclyde region, May-Jun 1992, Scotland. All the bold-formatted segments are part of the thread which refers to the teachers’ ballot for strike. However, one can observe that the segments of articles 17311 and 20570 are not relevant to that thread.
ARTICLE NO: 14055
TITLE: Tompson's sister tells murder trial of his dying words
DATE: 11/04/1992

FIGHTING back tears, the sister of Arthur Thompson Jr told the High Court in Glasgow yesterday how she ran into the street and saw her brother walking towards her, his arms outstretched, saying: “I have been shot, hen. I am going to collapse.”

Miss Tracy Thompson, 26, was giving evidence in the trial in which Mr Paul Ferris pleads not guilty to murdering Mr Thompson outside his home in Provanmill Road, Glasgow, on August 17 last year while acting with two other men who are now dead.

Mr Ferris also denies attempting to murder Mr Thompson, Sr, and attempting to murder a man in Rutherglen, attempting to pervert the course of justice, conspiracy to assault, illegal possession of drugs and firearms and breach of the Bail Act. He has lodged a number of special defences of alibi and incrimination.

Miss Thompson said that her brother had just left her parents’ house at No. 176 to go to his own house at No. 186 and she had walked him to the door. Moments later she heard noises.

“I thought it was a car back-firing and then I realised it was a gun going off,” she told the Advocate-depute, Mr Roderick Macdonald, QC.

She heard the noises about four or five times and said she then started panicking.

She went on: “I ran out into the street. I found Arthur with his arms outstretched coming towards me. There was no mark on him, there was no blood whatever, no mark, and I said: ‘What’s happened?’ He said: ‘I have been shot, hen. I am going to collapse.’ I grabbed him and he went down on his knees”...

ARTICLE NO: 17332
TITLE: Crown puts its case
DATE: 22/5/1992

THE Crown yesterday concluded its case on the 40th day in the Arthur Thompson, Jr, murder trial at the High Court in Glasgow.

The Judge, Lord McCluskey, heard legal submissions from both defence counsel and the Advocate-depute outwith the presence of the jury. The defence case is expected to open on Tuesday.

Mr Paul Ferris, 28, has pleaded not guilty to the murder of Arthur Thompson, Jr, outside his home in Provanmill Road, Glasgow, in August last year while acting with Robert Hanlon and Joseph Glover, who are now dead.

He also denies the attempted murder of Mr Arthur Thompson, Sr, the attempted murder of Mr Robert Johnstone, shooting Mr William Gillen in the legs to his severe injury and threatening to murder him, attempting to pervert the course of justice, conspiracy to assault, illegal possession of drugs and firearms and breach of the Bail Act. He has lodged a number of special defences of alibi and incrimination.

Earlier, Mr Donald Findlay, QC, for the defence, ended his address after speaking for six hours.

He is defending Mr Paul Ferris, 28, who denies murdering Mr Arthur Thompson, Jr, outside his home in Glasgow last August, while acting with Mr Robert Glover and Mr Joseph Hanlon, who are now dead.

He also pleads not guilty to the attempted murder of Mr Arthur Thompson, Sr, shooting William Gillen in the legs in a lay-by on the Glasgow-Kilmarnock road, illegal possession of drugs and a weapon, and breach of the Bail Act. He has lodged a number of special defences of alibi and incrimination.

Mr Findlay claimed that the failure of the Crown to call a key witness left a gap too big to ignore. The key witness was Thomas McGraw, alleged to have been the instigator of William Gillen being shot in the legs, but Mr McGraw had not given evidence, he told the jury.

He said Mr Gillen had claimed Mr McGraw was there on the night of the shooting and had ordered it...

ARTICLE NO: 22688
TITLE: Thompson trial jury set to retire today
DATE: 11/6/1992

THE Jury will retire today to consider its verdict in the Thompson murder trial as it goes into its 53rd day in the High Court in Glasgow.

The trial became the longest murder trial in Scottish legal history last week, beating the previous record of 47 days.

Lord McCluskey, the trial Judge, began his charge yesterday afternoon, and spoke for an hour before adjourning for the day. The jury consists of seven men and seven women. One juror was excused in the second week for medical reasons.

Mr Donald Findlay, QC, for the defence, ended his address after speaking for six hours.

He is defending Mr Paul Ferris, 28, who denies murdering Mr Arthur Thompson, Jr, outside his home in Glasgow last August, while acting with Mr Robert Glover and Mr Joseph Hanlon, who are now dead.

He also pleads not guilty to the attempted murder of Mr Arthur Thompson, Sr, shooting William Gillen in the legs in a lay-by on the Glasgow-Kilmarnock road, illegal possession of drugs and a weapon, and breach of the Bail Act. He has lodged a number of special defences of alibi and incrimination.

Mr Findlay claimed that the failure of the Crown to call a key witness left a gap too big to ignore. The key witness was Thomas McGraw, alleged to have been the instigator of William Gillen being shot in the legs, but Mr McGraw had not given evidence, he told the jury.

He said Mr Gillen had claimed Mr McGraw was there on the night of the shooting and had ordered it...

Table 5.7: Three articles relevant to Paul Ferris’ trial, Apr-May-Jun 1992, Scotland. Linking of summarizing paragraphs: observe the relationship between all the bold-formatted segments.
CHAPTER 5. EVALUATION

ARTICLE NO: 4854
TITLE: Steetley bids back in the UK
DATE: 13/02/1992

THE EUROPEAN Commission has agreed to let the UK competition authorities consider the proposed joint venture between Steetley and Tarmac.

Merger proposals above a certain size are automatically considered by the Commission now and Trade and Industry Secretary Peter Lilley requested that the matter be referred back to the UK. Mr Lilley argued that the merger would only affect the UK market and so was not relevant to other EC countries and the Commission agreed to this view. It is the first time a merger situation has been referred back to a Community country.

Now the proposed deal will be considered by the Office of Fair Trading which may recommend a reference to the Monopolies Commission. The Government is concerned for competition in the brick and clay roofing tile sectors.

<u>HE</u>

The enforced delay is playing into Steetley’s hands in respect of Redland’s £630m offer for Steetley as it gives more time for a defence to be prepared and for the brick market to build on the anecdotal indications that the market could be moving off the floor.

---

ARTICLE NO: 16682
TITLE: Thompson murder trial told how Barlinnie chief warned supergrass
DATE: 29/04/1992

---

A CROWN Office letter branding Mr Paul Ferris, the accused in the Thompson murder trial, a paid police informant for a number of years was declared a fake in the High Court in Glasgow yesterday by the offi-cial alleged to have written it.

Mr Findlay: A letter such as this conveyed around Glasgow might in the eyes of some people put Mr Ferris in a particular light, as a tout or grass?

Witness: Yes. Mr Paul Ferris, 28, has pled not guilty to having on May 26 last year while acting with Mr Robert Glover and Mr Joseph Hanlon, who are now dead.

He denies a charge that he murdered Mr Arthur Thompson, Jr, while acting with Mr Glover and Mr Joseph Hanlon, also dead. He also denies attempting to murder Mr Thompson Sr, attempting to murder a man in Rathrogen, attempting to pervert the course of justice, conspiracy to assassinate, illegal possession of drugs and $9 reams and breach of the Bail Act. He has lodged a number of special defences of alibi and inexcitation.

Mr Paul Hamilton spoke earlier of going to a football club dance in Kirkintilloch in April or May last year.

<pp>Redland argues that Steetley’s claim to have 150 million tonnes of reserves with its French Gobitta acquisition is misleading given that only 37 tonnes has planning consent for extraction and that in addition a dispute has arisen with the vendor over option rights on some of the land.

---

ARTICLE NO: 6028
TITLE: Redland now clear to pursue Steetley
DATE: 04/03/1992

... The Regiate-based company also said it would sell Steetley’s clay reserves at the latter two sites and at Walley’s pit in Staffordshire if required by a purchaser.

The undertakings were given by Redland as a result of discussions with the director general of fair trading, concerned about reduced competition in the UK clay tile and brick market. The Secretary of State announced on February 18 that the proposed acquisition by Redland of Steetley would be referred to the Monopolies and Mergers Commission unless suitable undertakings could be obtained.

Steetley’s preferred option to merge some of its operations with Tarmac in a new company was abandoned after the merger was referred to the Monopolies and Mergers Commission and no undertakings could be given without “seriously” undermining the industrial logic and rationalisation benefits of the venture.

Steetley is expected to offer its final defense tomorrow which will focus largely on asset valuations.

<pp>Mr Kerr was asked yesterday by Advocate-depute Iain Bonomy if he had any information about Woodman’s past history. “Did you know that he had on several occasions given informations about what he heard in prison which resulted in convictions being secured against other prisoners”...

---

Table 5.8: Examples of threads. 1st row: the story of merging between Tarmac and Steetley companies, thread of 1st row: the role of Monopolies and Mergers Commission of EC (observe the relationship between the bold-formatted segments). 2nd row: the story of Paul Ferris’ trial, thread of 2nd row: not clear (observe the relationship between the italic-formatted segments).
5.5 Overall evaluation

In addition to segment and thread evaluation, the overall performance of the NHS system was tested in terms of its usability. The NHS system has been designed in such a way that a user spends less time to fulfill her information need compared to classical IR system, using newspaper domain as the data source. Thus, user experiments were organized on the basis of comparing the performance of a classical IR system to the NHS one. The evaluation process is described as follows:\(^1\):

1. Two systems were used: the IR system and the NHS system. The former was the same as the NHS but without providing the threads interface.

2. Ten users were divided into two groups of five: IR group and NHS group. Users from the IR group (IR users) started using the IR system, whereas users from the NHS group (NHS users) started using the NHS system.

3. Each one of the five queries that were presented in the previous section was given to a pair of users, one user from the IR group and one from the NHS group. In response to that query, a set of articles was retrieved by each one of the two systems.

4. Every user wrote down a brief summary of the news story that the articles were about. However, while IR users had access to the whole set of retrieved articles, ordered by likely relevance, NHS users were restricted to use only the articles that were included in the thread interface.

5. Every user answered six questions relevant to the news story. Half of the questions had their answer in a thread, while the other half had their answers out of the threads. By this combination, the possibility of answering all the questions using for example only the thread interface of NHS system, and thus biasing the results, was prevented. NHS users were allowed to use the full facilities of the NHS system, that is the chronologically ordered list of retrieved articles and the thread interface. IR users had again access to the whole set of retrieved articles, ordered by likely relevance.

6. Finally, the two groups changed roles: the NHS users became IR users and were given the IR system, whereas the IR users became NHS users and were given the NHS system. In both occasions, the users were asked to spend some time in browsing articles and express their opinion about the usefulness of the thread interface. Specifically, the new IR users (previous NHS ones) were asked if the thread interface helped them in understanding the story and answering the questions easily. The new NHS users (previous IR ones) were asked if the thread interface could have helped them in understanding the story and answering the questions easier.

\(^1\)the reader should refer to the appendix whenever details concerning the interface of the NHS system are presented
It must be noticed that the users were grouped in pairs according to their academic background, their knowledge of English language and their experience with computer systems, in order to ensure fair results. All users were undergraduate or postgraduate students who were not computer experts. However they were all familiar with WWW search engines. Every user had a maximum time period of 30 minutes to write the summary and answer the questions. During that time, discussions with the experimenter were not allowed. However, all the users’ actions were logged automatically by the system. After the period of answering questions, the users were encouraged to express their comments, especially in the 6th step of the evaluation process.

The results of the experiment are presented in figure 5.2. The figure indicates that the NHS users answered the questions faster and more accurately than the IR users. Moreover, all the users found the NHS system more effective in terms of its help for understanding the news story and answering the questions.

Comparing the brief summaries that the users wrote down, one can observe that the NHS users, despite the fact that they were not allowed to read all the retrieved articles but only the ones that were contained in the detected threads, could achieve the same level of understanding the basic aspects of the stories with the IR users. This happened because summarizing paragraphs, which are repeated in many articles usually at their beginning or their end, had been identified as similar segments and they had been linked.

The log files of users’ actions as well as the discussion that took place between the experimenter and each individual user provided interesting information:

<table>
<thead>
<tr>
<th></th>
<th>IR</th>
<th>NHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIME (min)</td>
<td>19</td>
<td>14</td>
</tr>
<tr>
<td>FAULTS</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MARK NHS</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

QUERY 1 (Q1)

<table>
<thead>
<tr>
<th></th>
<th>IR</th>
<th>NHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIME (min)</td>
<td>30</td>
<td>27</td>
</tr>
<tr>
<td>FAULTS</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MARK NHS</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

Figure 5.2: Overall evaluation results for the NHS system: One query per two users (IR user, NHS user)
Table 5.9: Threads that were detected for the news story of Julie Ward’s murder in Kenya (Feb 1992), using 50 and 100 articles respectively: notice that with 100 articles, one more thread is detected which is irrelevant. This thread is related to Judith Ward, an IRA member, who was accused of bombing an Army coach using nitro-glycerine. The two women’s common surname was the reason of having these non relevant articles.

1. The users browsed the articles mainly from the temporal layout and not by following the usual hypertext links.

2. They also claimed that they found helpful the easy detection of recent articles or old articles relevant to a story. The former ones helped them to identify easily the conclusion of the story, while the latter ones provided the basic source for writing down the summary.

3. The users did not come across non-relevant threads when cut-offs in the number of retrieved articles were used. Without using cut-offs, non-relevant threads of segments of non-relevant articles that happened to form stories, non-related to the main news story, were scarcely appeared. One such occasion is presented in table 5.9.

4. The majority of the users were against narrow-specified queries. They preferred using short queries and having the system classified the results into various groups of topics.

5. Finally, all users had the impression that the links had been included in the articles manually and asked about the time requirements of such a process.

It must be noticed that the whole evaluation process was based on queries that had news stories with threads as a result. The NHS system takes advantage of such kind of queries and generates threads, where possible. If threads do not exist for a specific query, then the NHS system provides the user only with a set of relevant articles and does not present threads. In that case, the NHS system is nothing more than an IR engine. The evaluation
was performed in such a way in order to test the NHS system as an extension to a normal IR system, when the latter needs to deal with newspaper archives.

5.6 Conclusion

In this chapter, the effectiveness of the NHS system was evaluated. The evaluation was based on metrics as well as on user tests and was performed with a set of five queries (news stories).

Segment evaluation was performed in order to evaluate the quality of the decomposition process. Segmented articles were given to users, who were in general satisfied with the results, although sometimes the boundaries of the segments were not well-defined.

Compactness and stratum were the two metrics that examined the structure of the resulted hypertext for its linear ordering and its complexity. The results indicated the presence of cyclic graphs with gaps which provide the user with distinguishable and well-separated sets of related segments. These sets reflect the notion of threads.

Users evaluated the threads that the NHS system detected for news stories, by reading the linked segments and marking the linking quality. The linking was characterized as successful by the majority of the users.

Finally, the NHS system was compared with a normal IR system in terms of the time that a user spends in order to understand the content of a news story and answer correctly some questions relevant to it. The results showed that the users of the NHS system could answer questions relevant to a news story faster and more accurate than the users of the IR system.

The evaluation results give the indication that the NHS system has fair success in meeting the requirements of information retrieval using newspaper domain.
Chapter 6

Further work

6.1 Introduction

The evaluation results of many aspects of the NHS system (decomposition, thread construction, overall performance) showed its viability as a hypertext information retrieval system which deals with the newspaper domain. Its automatically constructed hypertext can provide an effective way of browsing news articles.

However, there are several issues that need to be considered, both in the implementation level and the evaluation process, for improving the NHS system and testing its performance. The following sections discuss these issues in relation to the limitations of the NHS system.

6.2 Implementation level

- **Text decomposition**
  Since the decomposition of articles into segments is crucial for subtopic detection and thread construction, other alternative approaches should be tested. The method that was used in the NHS system had fair success in general. Its major drawback is that the boundaries of the segments are sometimes over-extended or incomplete, because the NHS decomposition engine does not verify the decision for segment detection by looking also at small text units, like for example sentences.

TextTiling system (Hearst and Plaunt, 1993; Hearst, 1994) is a research software tool that partitions text documents into coherent multi-paragraph units. TextTiling uses patterns of lexical connectivity in order to find out sub-discussions within the document. Similarity values between pairs of adjacent text blocks are computed and the resulting sequence of these values is examined. Peak values indicate the presence of a subtopic while low values indicate the presence of a potential boundary between subtopics. The text blocks that are used are smaller than paragraphs.
The TestTiling approach can be exploited by the NHS system for more accurate text decomposition.

- **Relevance feedback**
  Effective subtopic identification depends on the number of relevant articles which are retrieved. The more relevant relevant are retrieved, the more subtopics are detected. One way to increase the number of relevant articles that are retrieved (high recall) is to perform relevance feedback, by giving the user the opportunity to indicate articles that she considers to be relevant to the news story after an initial search. Then the NHS system can use these articles as a source for new query terms that will added to the previous ones so that the final search is performed.

- **Clustering**
  In the NHS system, the clustering level for the single-link is selected by examining the changes in the number of clusters for various similarity levels. A more accurate process would be to observe the changes in the graph structure that is derived from the maximum spanning tree (see chapter 4) while the similarity level changes, too. Euclidean distances between matrices which represent the various graph structures can be used to measure the degree of change between these graph structures.

- **Handling irrelevant documents**
  The story chain may contain a number of irrelevant documents. In case that this number is big, the user will be annoyed during the process of browsing the story. There is strong evidence that irrelevant documents form small threads within the main story chain and thus they can be eliminated (an example was presented in section 5.5). The idea has been applied successfully in (Hearst and Pedersen, 1996), too.

- **Handling time gaps**
  In case of queries with a general subject, there may be large time gaps for the story that the retrieved articles are related to. An example could be a query asking for articles relevant to the terms “elections” and “United Kingdom”. In that case, a separation of the whole story to many small ones that correspond to a time period without large time gaps might be considered.

- **Thread surrogates**
  Thread surrogates are necessary for understanding the story evolution. In the current prototype of the NHS system, the first sentences are used to give the user an indication of its content. However, other approaches involving keyword extraction, theme identification and summarization should be also considered. For a recent work in indicative summarization (that is summaries which help the users to decide whether a document is about the subject that they search for), which is also query-oriented, one can refer to (Tombros, 1997).
6.3 Evaluation process

- Evaluating the retrieval engine
  Relevance judgements were not available for the “Herald” collection. Thus, there is no indication of the retrieval effectiveness of the SMART-based retrieval engine. For this reason, user tests are also needed in order to estimate its retrieval performance. The “Herald” collection was preferred against other newspaper collection with relevance judgements, like the “Wall Street Journal” and the “Financial Times” collections, because it covers a wide range of news, as opposed to the other two.

- Evaluating the threads
  The thread evaluation was “precision-oriented”. Users evaluated the quality of the detected threads but they did not indicate threads that were not detected. The users tests should be extended in order to include the “recall” aspect of linking: the proportion of good links that were detected.

- Evaluating the clusters
  The quality of clustering was tested indirectly by evaluating the quality of the threads. However, cluster validity techniques (see section 5.2) can provide an indication of the existence or not of clustering tendency within the sets of segments and thus can help to confirm the results from the user experiments.
Chapter 7

Conclusions

This thesis presented a methodology for the automatic construction of links for news hypertext which is tailored to the domain of newspaper archives.

Basic issues of information retrieval and hypertext were discussed. Information retrieval uses direct search techniques for accessing the content of documents. Content representation is performed using index terms that are derived from the documents via automatic text analysis and weighting techniques. On the other hand, hypertext provides access to the various kinds of associations that may exist between documents or their parts. Navigation and browsing help the user to discover such associations and thus to create a dynamic information path. Hypertext information retrieval systems combine hypertext and information retrieval capabilities by providing retrieval techniques that include direct search as well as navigation and browsing. The incorporation of a mechanism for the automatic construction of hypertext into an IR system is quite important in order to exploit the advantages of the integration of hypertext with information retrieval.

Formal ways to capture the temporal aspects that characterize the newspaper domain, and which are often ignored by IR systems, were suggested. Aggregate links and thread links were used in order to describe the evolution of a news story. Aggregate links connect articles related to a story in a chronologically ordered chain. Thread links connect articles that totally or partially refer to a substory within the main story in some of their segments. Using such an approach, one can easily visualize the evolution of a story together with the evolution of its threads, by constructing temporal layouts.

The above model of threads and stories was used for the implementation of the NHS (News Hypertext System) system which automatically creates links for news hypertext. The NHS system retrieves news articles relevant to a query and presents a story with threads as a result. A story is created by linking the articles in a chronologically ordered chain. Threads are constructed by linking related segments of articles in the hope of capturing different substories. Threads also form chronologically ordered chains. Related segments are identified by applying clustering techniques to all the segments of the retrieved articles. The clustering process, which is the core of the hypertext construction, is
CHAPTER 7. CONCLUSIONS

The effectiveness of the NHS system was evaluated. The evaluation was based on metrics as well as on user tests and was performed with a set of five queries (news stories). Segment evaluation was performed in order to evaluate the quality of the decomposition process. Segmented articles were given to users, who were in general satisfied with the results even though sometimes not well-defined boundaries were indicated. Compactness and stratum were the two metrics that examined the structure of the resulted hypertext for its linear ordering and its complexity. The results indicated the presence of cyclic graphs with gaps which provide the user with distinguishable and well-separated sets of related segments. These sets reflect the notion of threads. Users evaluated the threads that were detected for news stories, by reading the linked segments and marking their quality. The linking was characterized as successful by the majority of the users. Finally, the NHS system was compared with a normal IR system in terms of the time that a user spends in order to understand the content of a news story and answer correctly some questions relevant to it. The results showed that the users of the NHS system could answer questions relevant to a news story faster and more accurate than the users of the IR system.

The evaluation results give the indication that the NHS system may have fair success in meeting the requirements of information retrieval using newspaper domain. However, many issues need to be further considered, both in the implementation level and the evaluation process. Among them, other decomposition techniques should be tested, a relevance feedback mechanism should be provided for high recall and the “recall” aspect of linking should be incorporated into the evaluation procedure.
Bibliography


Appendix A

The NHS prototype

Figure A.1: The main page of the NHS system: The query “Julie Ward Murder” has already been typed. The retrieved articles can be ordered by date of publication or by relevance.
Figure A.2: Retrieved articles from the NHS system: Observe that the articles have been ordered by date of publication.
Figure A.3: An example of an article which is part of a thread: A segment has been marked with the labels “START” and “END”. The link “LINK TO...” points to the next related segment of another article.
Nairobi, Friday. VULTURES perched in a remote corner of Kenya’s Masai Mara game reserve led a search party to the remains of murdered British tourist Julie Ward, a police witness told a Nairobi court today.

"When you see vultures around you know there must be an animal killed somewhere," Inspector George Odhiambo told the trial of two game rangers accused of killing Julie in 1998.

Odhiambo said he and a search party of game rangers spotted the vultures perched in a tree while searching a bushy, rocky area of the reserve.

"We then found two pieces of human jaw and later we found a left leg of a human being," Odhiambo said.

* ENTRY POINT *

----- START ---------------------------

----- START ---------------------------

Earlier today Julie’s father wound up two weeks of evidence by telling the court his daughter was probably killed on the same day he searched and missing remains were found, on 18 September, 1998.

Julie, 28, disappeared eight days earlier while on a camping holiday in the Masai Mara.

----- END ---------------------------

----- END ---------------------------

* LINK FROM *

----- END ---------------------------

* LINK TO *

"I’m free at last," John Ward said as he left the witness stand to join his wife Janet in the public benches.

Game rangers Jonas Tajeu Megino, 28, and Peter Metui Kirerir, 28, face the death penalty if convicted.

Figure A.4: The next segment related to the segment of the article in figure A.3. The segment has been marked again with the labels “START” and “END”. The link “LINK TO...” points to the next related segment of another article, while the link “LINK FROM...” points to the previous related segment.
Figure A.5: The temporal layout that the NHS system provides: The oval buttons represent articles whose segments are part of a thread. By pressing these buttons, the articles are presented to the user (see figure A.3). The color of the buttons represents the degree of relevance of their corresponding articles. Furthermore, surrogates for each thread (the first sentences of the first segment of the thread) are presented.
APPENDIX A. THE NHS PROTOTYPE

Figure A.6: The story evolution of a thread: Surrogates of the articles whose segments are part of a thread are presented (the first sentences of their related segments).
Appendix B

Sample of the evaluation instructions