NHS: A Tool for the Automatic Construction of News Hypertext

Theodore Dalamagas[*]
Computer Science Division
National Tech. Univ. of Athens
Zographou Campus, Greece, 15773
email: dalamag@dbnet.ece.ntua.gr

Abstract

The automatic construction of hypertext is an important part of the hypertext authoring process. This paper presents the NHS system, a system that automatically creates links for news hypertext which is tailored to the domain of newspaper archives. 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 automatically linking these segments with segments in subsequent articles. The evaluation of such an approach to the automatic construction of hypertext is finally discussed, in terms of its usability and the structural quality of resulting hypertext.

1 Introduction

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 proposed by Allan [Allan, 1996], 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 works on the automatic transformation of a well-structured document into hypertext, only structural links are identified. Rada [Rada, 1992] 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. Salton et al. [Salton and Buckley, 1989, Salton et al., 1993, Salton et al., 1994a] use the information provided by the computation of the similarity between fragments of documents in order to identify content links. Smeaton et al. in [Smeaton and Morrissey, 1995] use also the similarity between document fragments but they selectively add links depending on how this influences values of various topology measures. Agosti et al. [Agosti et al., 1996] propose a conceptual architecture for information retrieval systems, structured on three levels: documents, index terms and concepts. Based on this architecture, they present a methodology for the automatic construction of links between objects within each of these levels and between levels.

This paper describes the NHS system which automatically creates news hypertext for the domain of newspaper archives. The work was initiated in [Dalamagas and Dunlop, 1997], where a theoretical framework of a methodology for the automatic construction of news hypertext was discussed. This paper concentrates mainly on implementation and evaluation issues, which have not been sufficiently addressed in [Dalamagas and Dunlop, 1997]. The following section discusses formal aspects of news hypertext and modeling issues. The third section describes the NHS system and the method that is used in order to automatically construct hypertext for a set of retrieved articles relevant to a query inserted by the user. Evaluation issues follow and, finally, directions of further work and other conclusions are discussed.

[*] The work was carried out while the author was doing his MSc in Advanced Information Systems in the Computing Science Department of Glasgow University.

2 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 is called *thread chain*, or simply *thread*.

As an example, consider the recent news story of TWA air crash. The story evolved through the publication of a great number of related articles over a long period of time. The evolution of the story started from the initial reports of the accident and continued with “missile” theories, “bomb” theories, “missile” theories (again)[1], compensation issues, etc. The sequence of articles that refer to “missile” theories is a thread of the main story of the TWA air crash. The evolution of the story is presented in figure 1.

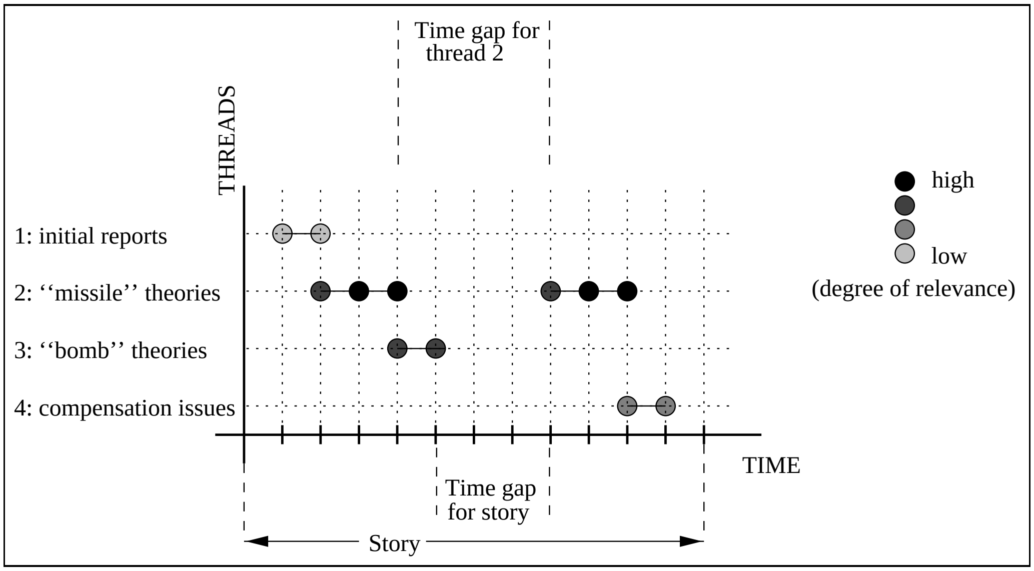


Figure 1: Temporal layout of story evolution

Figure 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 1. The figure 1 is actually the *temporal layout*[2] of the story and it offers a simple but intuitive method to visualize the evolution of the story together with the evolution of threads, providing temporal semantics.

The object model [Rumbaugh et al., 1991] can be used to describe formally the suggested scheme, as depicted in figure 2. A story consists of articles and A-links which connect them. Similarly, a thread consists of articles’ *segments*

[1] “missile” and “bomb” theories refer to the cause of the accident

[2] the temporal layouts are used to describe action scenarios in multimedia applications, e.g. video and audio playback during a period of time

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 1, which are not explicitly modeled 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.

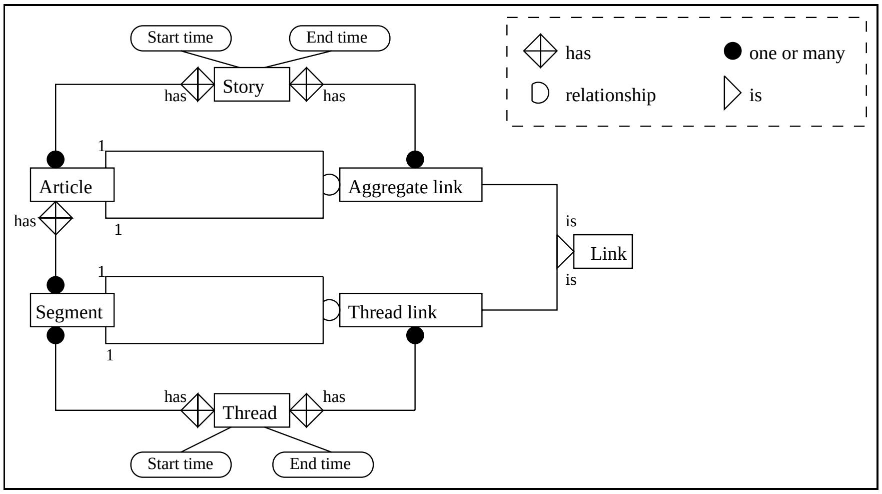


Figure 2: Object model

The above model will be used in the following sections as the framework for the development of methods for the automatic construction of news hypertext. A conceptual model for navigating and browsing among different IR objects has been also used in [Agosti et al., 1996] (see section 1). 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 which are presented in the following sections, in contrast to the manual or thesaurus-based construction of concepts' set [Agosti et al., 1996].

3 NHS: A system that automatically creates news hypertext

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 paper 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 3.

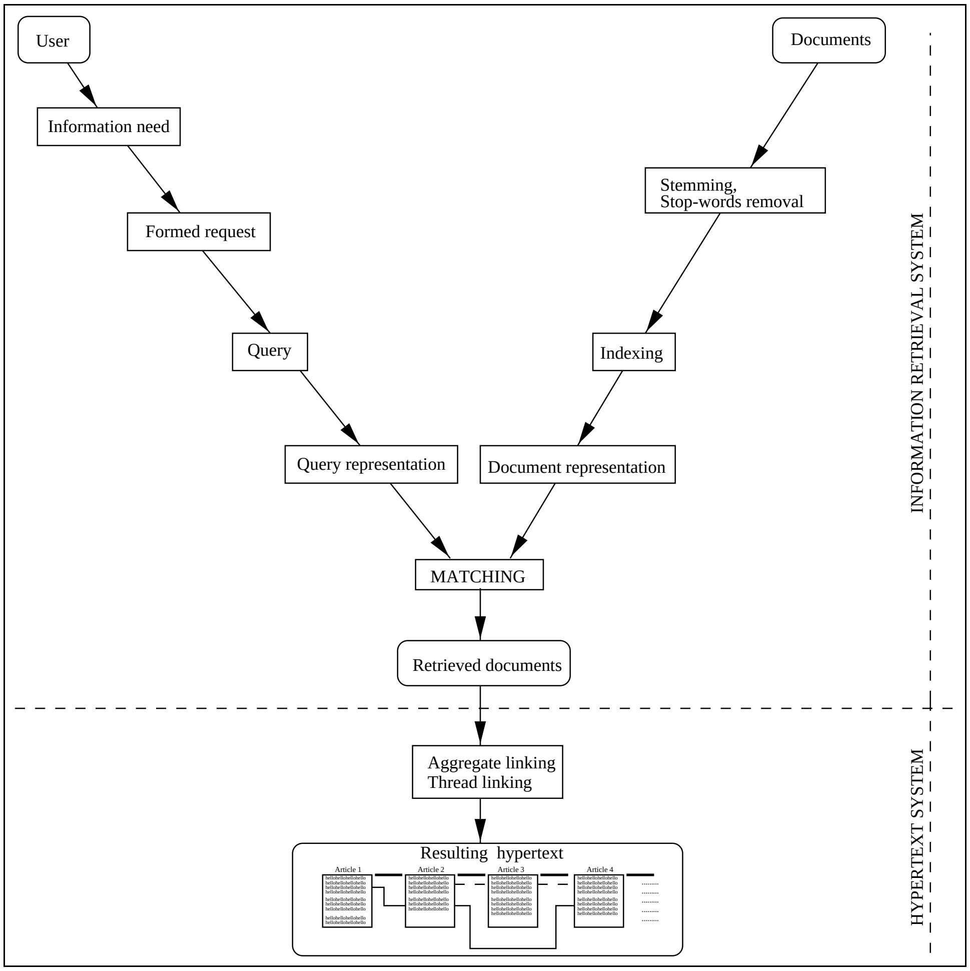


Figure 3: Architecture of the NHS system

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, which has been used in the NHS system, is summarized as follows:

- Decompose all the articles of the collection into segments, prior to any usage of the NHS 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.

3.1 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 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.

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).

3.2 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 is based on the one explored by Salton et al. and is described in the following section [Salton et al., 1994b, Salton et al., 1995].

3.2.1 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 4a. 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 starts with the construction of the paragraph relationship map of the document that needs to be decomposed into segments (see figure 4a). Then, all *low-similarity* links (that is links of the map that correspond to similarity values which are below a predetermined threshold) as well as all *long-distance* links (that is links of the map spanning more than a predetermined number of adjacent paragraphs) are dropped (see figures 4b and 4c). 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 4c, 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[3] of the paragraph relationship map (graph). The main characteristic of these connected components is the lack of long-distance links.

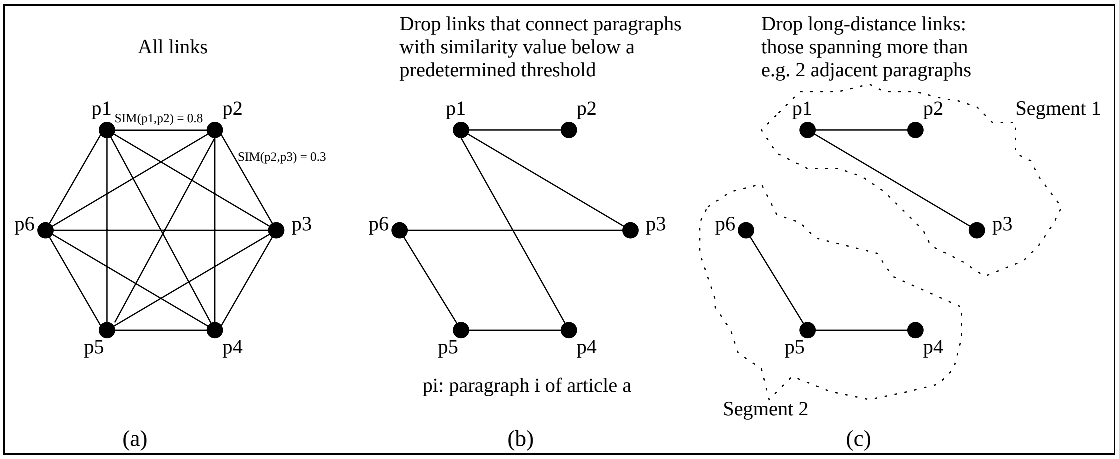

All links
Drop links that connect paragraphs
with similarity value below a
predetermined threshold
Drop long-distance links:
those spanning more than
e.g. 2 adjacent paragraphs
p1
p2
SIM(p1,p2) = 0.8
SIM(p2,p3) = 0.3
p3
p4
p5
p6
Segment 1
Segment 2
pi: paragraph i of article a
(a)
(b)
(c)

Figure 4: Simplification of paragraph relationship map for segment detection

3.2.2 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. Using the SMART system, the paragraphs of an article are identified and indexed as separate documents and similarity values are computed for each pair of paragraphs. 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.

2. Similarity values between long-distant paragraphs are not taken into account, too. Two paragraphs are considered to be long-distant if they have more than 1 paragraphs to exist between them within the article. 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 relationship map of the specific article. These values are examined so that chains of adjacent linked paragraphs are identified within the article. These chains are a first indication of possible segment formation. One example of such a chain is the set of $p4-p5-p6$ paragraphs in the map of figure 4c. The initial chains may be extended if links to other paragraphs are found at the beginning or the end of a chain, based on the fact that paragraphs which are related to the same one give a possible indication that all of them are somehow related. This is depicted clearly in figure 5. 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:

$$a_i \longrightarrow (S_{i1}, S_{i2}, \ldots, S_{in_i}) \tag{1}$$

[3] 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

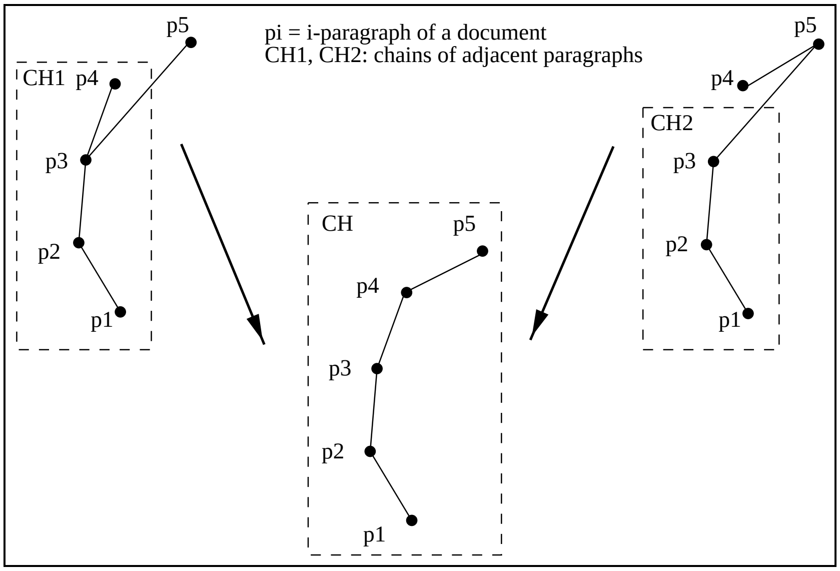

pi = i-paragraph of a document
CH1, CH2: chains of adjacent paragraphs
CH1
CH2
CH
p1
p2
p3
p4
p5

Figure 5: Extending chains of adjacent paragraphs: Chain CH1 initially consists of paragraphs $p1, p2, p3, p4$. Chain CH1 is extended to chain CH in order to include also $p5$. Similarly, chain CH2 is extended to chain CH in order to include also $p4, p5$.

After the segments of each article have been identified, they are indexed as separated documents using the SMART system. The indexes are needed for the computation of similarity values, as it will be explained in the next section.

3.3 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 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. The n segments of the retrieved articles form a graph G with n vertices $\in V$ and $n(n-1)/2$ weighted edges $\in E$. The weight of an edge corresponds to the similarity value between the vertices (segments) that this edge connects. It has been shown [Gower and Ross, 1969] that a MST contains all the information needed in order to perform single link clustering: 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.

Three problems 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

3.3.1 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).

3.3.2 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 $S = (S_1, S_2, \ldots, S_k)$ 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(O1, O2)*: unites the two sets S_{O1}, S_{O2} that contain $O1$ and $O2$, 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 FIND-SET(v) $\neq$ FINDSET(u) then UNION(v, u)

3.3.3 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 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 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 $0.48, 0.46$ give as another possible final selection the level 0.46 (1 refinement step: 0.46). Because the level 0.56 comes from more refinement steps than the other two levels, it will be used finally as the clustering level. The above methodology is described analytically in [Dalamagas, 1997].

The performance of single link clustering does not impose great overhead on the system. Clustering of 200 segments for all similarity levels between 0.8 and 0.2 with step 0.02 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.

CLUSTERING LEVEL	NUMBER OF EDGES	NUMBER OF CLUSTERS
. . .	. . .	. . .
0.62	7	4
0.6	9	6
0.58	10	6
0.56	10	6
0.54	12	8
0.52	14	10
0.5	16	9
0.48	18	10
0.46	22	10
0.44	24	9
0.42	27	8
. . .	. . .	. . .

Table 1: Refinements of a single link clustering process

The automatically created hypertext has the form that is depicted in the example of figure 6. Despite the fact that the articles in figure 6 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 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 paper 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.

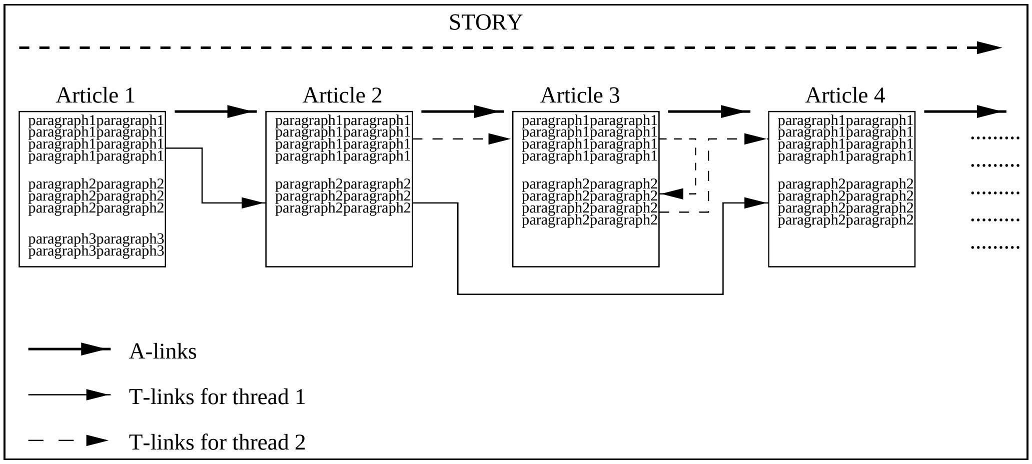


Figure 6: News hypertext

3.4 NHS prototype

A working prototype of the NHS system has been implemented[4]. Its structure extents the three-level architecture that was presented in the previous paragraphs by adding a WWW interface 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 7.

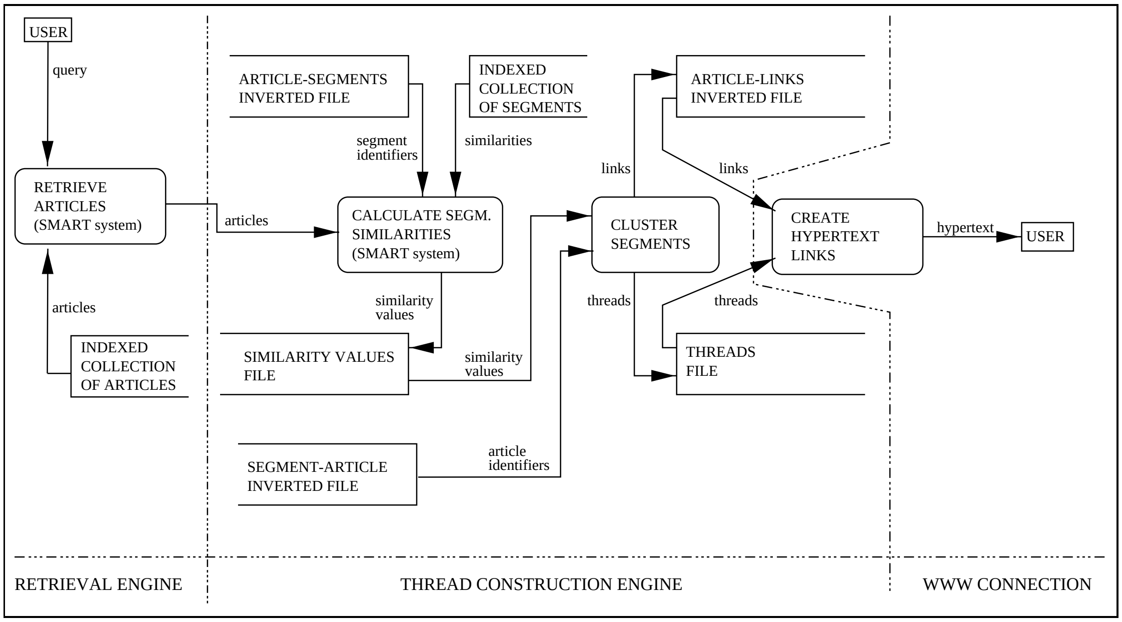
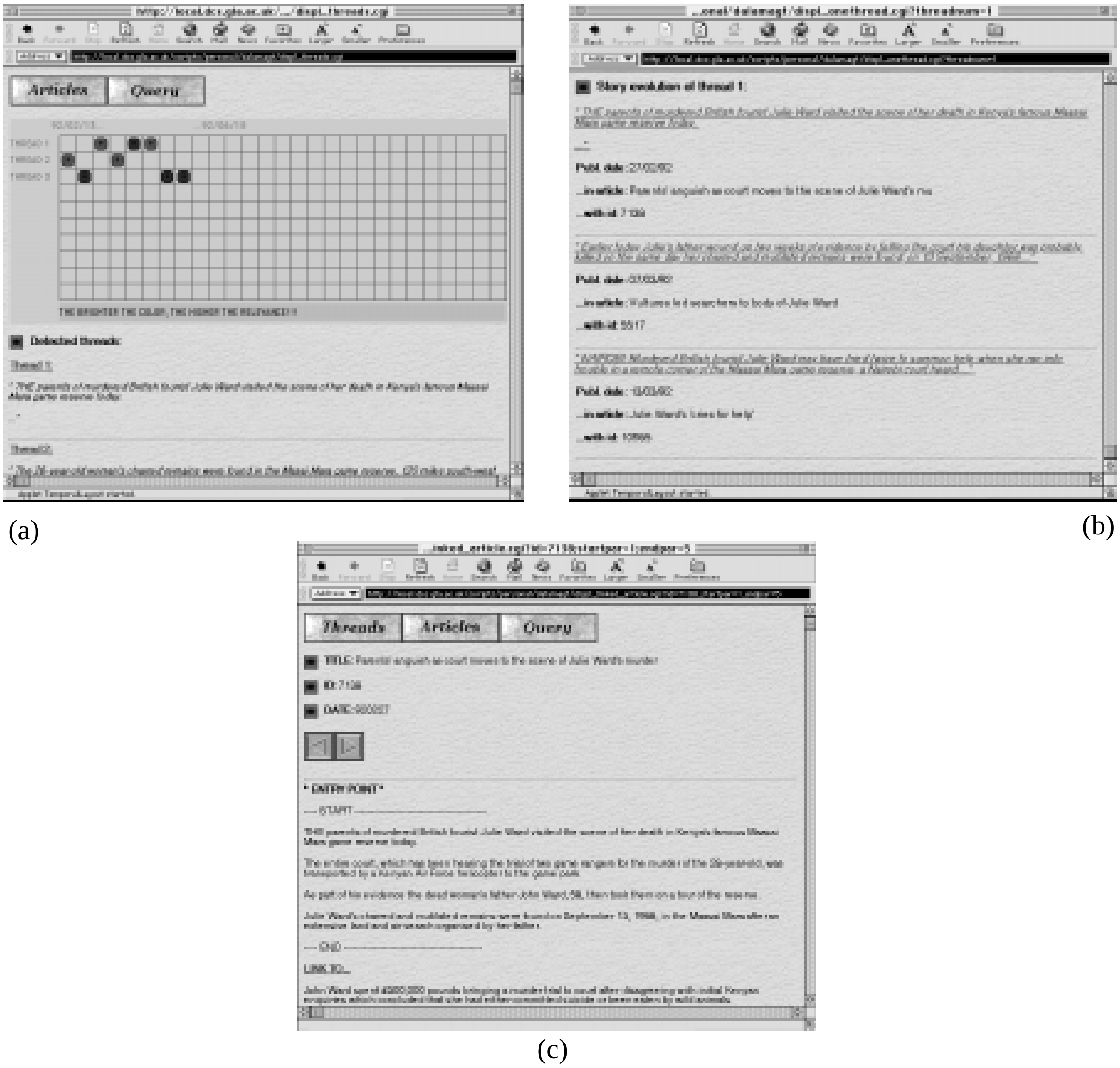


Figure 7: The NHS system: data flow

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 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 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 7: WWW connection).

In figure 8a the temporal layout that the NHS system provides is depicted. The oval buttons represent articles whose segments are part of a thread. By pressing one of these buttons, the corresponding article is presented to the user. 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. In figure 8b the story evolution of a thread is presented. Finally, figure 8c depicts 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.

[4] PERL script language, Java, C language and CGI scripts are the software tools that were used



THE BRIGHTER THE COLOR, THE HIGHER THE RELEVANCE!!!

Detected threads:

Thread 1:

* THE parents of murdered British tourist Julie Ward visited the scene of her death in Kenya's famous Masai Mara game reserve today.

...*

Thread2:

* The 28-year-old woman's charred remains were found in the Masai Mara game reserve, 120 miles south-west

Applet TemporalLayout started

Story evolution of thread 1:

* THE parents of murdered British tourist Julie Ward visited the scene of her death in Kenya's famous Masai Mara game reserve today.

...*

Publ. date: 27/02/92

...in article: Parents' anguish as court moves to the scene of Julie Ward's mu

...with id: 7138

* Earlier today Julie's father wound up ten weeks of evidence by telling the court his daughter was probably killed on the same day her charred and mutilated remains were found on 13 September, 1988...*

Publ. date: 27/03/92

...in article: Vultures led searchers to body of Julie Ward

...with id: 5517

* NAIROBI: Murdered British tourist Julie Ward may have [illegible] when she ran into trouble in a remote corner of the Masai Mara game reserve, a Nairobi court heard...*

Publ. date: 12/03/92

...in article: Julie Ward's 'cries for help'

...with id: 10566

Applet TemporalLayout started

TITLE: Parents' anguish as court moves to the scene of Julie Ward's murder

ID:7138

DATE:920227

* ENTRY POINT *

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

THE parents of murdered British tourist Julie Ward visited the scene of her death in Kenya's famous Masai Mara game reserve today.

The entire court, which has been hearing the trial of two game rangers for the murder of the 28-year-old, was transported by a Kenyan Air Force helicopter to the game park.

As part of his evidence the dead woman's father John Ward, 58, then took them on a tour of the reserve.

Julie Ward's charred and mutilated remains were found on September 13, 1988, in the Masai Mara after an extensive land and air search organised by her father.

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

LINK TO...

John Ward spent £500,000 pounds bringing a murder trial to court after disagreeing with initial Kenyan enquiries which concluded that she had either committed suicide or been eaten by wild animals.

Figure 8: (a) The temporal layout of the NHS system. (b) The story evolution of a thread: surrogates of the articles whose segments are part of a thread are presented. (c) An example of an article which is part of a thread.

4 Evaluation

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*
  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. The mathematical foundations of structural analysis of hypertext can be found in [Botafogo et al., 1992].

- *Hypertext quality*
  The constructed threads should satisfy the users, in the sense that they provide them with a set of related segments of various articles which refer to the same topic. Hypertext quality is mainly affected by the decomposition process that is used in order to identify segments within an article and by the clustering techniques.

- *System usability*
  The effectiveness of the NHS should be tested from the user's point of view. 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 as well as the accuracy of the user understanding.

Based on the above remarks, a small-scale evaluation procedure was performed in the following three parts: segment evaluation, thread evaluation and overall evaluation.

4.1 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 reading decomposed articles and answering questions. As one can see in table 2, 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).

QUESTIONS	YES	NO
Do the sets of paragraphs that are marked form a real segment?	9	1
Are there more segments that were not found?	3	7
Do the sets of paragraphs that are supposed to form a segment contain more (or less) paragraphs that are really needed?	5	5
Are you satisfied with the results?	7	3

Table 2: Segment evaluation results from user experiments

4.2 Thread evaluation

4.2.1 Hypertext metrics

Five queries that correspond to five different long news stories were randomly picked up. The selected queries cover a wide range of news (social, economic, international etc), as one can see in table 3. 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 9. 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*, two measures that deal with the overall topology of the hypertext [Botafogo et al., 1992]. 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 ($\cong 0$). 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.

NUMBER	NEWS STORY	KEYWORDS	ARTICLES
1	The Ferris murder case	paul, ferris, murder	70
2	The negotiations between Arabs and Palestinians	arab, palestinian, peace, negotiation	80
3	The Ward murder case	julie, ward, murder	50
4	The story of merging between Tarmac and Steetley company	redland, braas, steetley, company, germany, tarmac	30
5	A teachers' strike	teacher, strike, educational, institute, scotland	100

Table 3: News stories

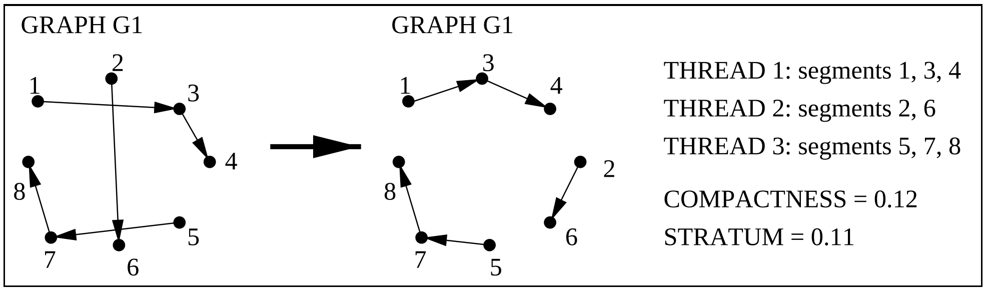


Figure 9: Expected form of news hypertext

The calculated values of compactness and stratum of the resulting hypertext for the five aforementioned queries are presented in table 4. The results confirm the expectation for the hypertext structure, giving an average of 0.14 for compactness and 0.15 for stratum.

4.2.2 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 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 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

NUMBER	COMPACTNESS	STRATUM	LINKS/NODES
1	0.13	0.14	10/14
2	0.12	0.025	13/19
3	0.12	0.11	5/8
4	0.17	0.25	2/4
5	0.16	0.21	9/12
AVG	0.14	0.15	-

Table 4: Compactness and stratum for the resulting hypertext of the five stories

U1	U2	U3	U4	U5	U6	U7	U8	U9	U10
2.4	2.5	3	2.8	3.2	2.8	3.5	3.5	3.3	3.3
Q1	**Q1**	**Q2**	**Q2**	**Q3**	**Q3**	**Q4**	**Q4**	**Q5**	**Q5**

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

contains, helped them to easily detect the false links and not take them into account while browsing.

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.

3. Non relevant threads, that is threads that link segments which belong to non relevant articles, are rare.

4.3 Overall evaluation

In addition to segment and thread evaluation, the overall performance of the NHS system was tested in terms of its usability, compared to a classical IR system.

Every set of articles relevant to each one of the news stories of table 3 were given to a pair of users, one using the NHS system and one using the IR system, which was the same as the NHS but without the threads interface. 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 wrote down a brief summary of the news story and answered some questions relevant to that story. Then all users 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.

The results of the experiment are presented in figure 10. The figure indicates that the NHS users answered the questions faster and more accurate than the IR users. Moreover, all the users found the NHS system more effective comparing to the IR system, 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.

It must be noticed that the 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.

The whole evaluation procedure has been described analytically in [Dalamagas, 1997].

5 Conclusions and further work

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

A formal way to capture the temporal aspects that characterize the newspaper domain was 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

	IR	NHS
TIME (min)	19	14
FAULTS	1	0
MARK NHS	3	3

QUERY 1 (Q1)

	IR	NHS
TIME (min)	30	27
FAULTS	1	0
MARK NHS	4	3

QUERY 2 (Q2)

	IR	NHS
TIME (min)	20	13
FAULTS	1	0
MARK NHS	4	3

QUERY 3 (Q3)

	IR	NHS
TIME (min)	25	19
FAULTS	1	0
MARK NHS	4	4

QUERY 4 (Q4)

	IR	NHS
TIME (min)	27	26
FAULTS	2	0
MARK NHS	3	4

QUERY 5 (Q5)

TIME (min): time to answer the questions and write the summary

FAULTS: number of wrong answers

MARK NHS: how the user marks the NHS system, compared to the IR system,
in terms of its help for understanding the news story
and answering the questions
Grade scale: 1, 2, 3, 4

Figure 10: Overall evaluation results for the NHS system: One query per two users (IR user, NHS user)

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 done on-the-fly and only for the set of the retrieved articles.

A recent proposal of integrating hypertext and information retrieval especially for the newspaper domain can be found in [Golovchinsky and Chignell, 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.

A small-scale evaluation of the NHS system was performed. The evaluation was based on metrics as well as on user tests and was performed with a set of queries (news stories). The results showed that the automatic construction of hypertext works well and users took advantage of it in order to fulfil their information need fast and accurately.

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.

- *Text decomposition*
  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. 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.

- *Evaluation*
  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. Moreover, 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. Finally, the evaluation procedure should be extended in larger scale.

Acknowledgement

I am grateful to Mark D. Dunlop for supervising this work. I also acknowledge the useful comments of Marcos Theophylactou and Yannis Theodoridis.

References

[Agosti, 1996] Agosti, M. (1996). An overview of hypertext. In Agosti, M. and Smeaton, A., editors, *Information Retrieval and Hypertext*. Kluwer Academic Publishers.

[Agosti et al., 1996] Agosti, M., Crestani, F., and Melucci, M. (1996). Design and implementation of a tool for the automatic construction of hypertext for information retrieval. *Information Processing and Management*, 32(4):459–476.

[Allan, 1995] Allan, J. (1995). *Automatic Hypertext Construction*. PhD thesis, Cornell University, Ithaca, New York. Also technical report TR95-1484.

[Allan, 1996] Allan, J. (1996). Automatic hypertext link typing. In *Proceedings of the ACM Hypertext'96 Conference*, pages 42–52, Washington, D.C., USA.

[Botafogo et al., 1992] Botafogo, R. A., Rivlin, E., and Schneiderman, B. (1992). Structural analysis of hypertexts: identifying hierarchies and useful metrics. *ACM Transactions on Information Systems*, 10(2):142–180.

[Buckley, 1985] Buckley, C. (1985). Implementation of the SMART information retrieval system. Technical Report TR85-686, Department of Computer Science, Cornell University, Ithaca, New York.

[Cormen et al., 1990] Cormen, T., Leiserson, C., and Rivest, R. (1990). *Introduction to algorithms*. MIT Press, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142.

[Cutting et al., 1992] Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W. (1992). Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318–329, Copenhagen, Denmark.

[Dalamagas, 1997] Dalamagas, T. (1997). Automatic construction of news hypertext. Master's thesis, Computing Science Department, University of Glasgow, Glasgow, Scotland.

[Dalamagas and Dunlop, 1997] Dalamagas, T. and Dunlop, M. D. (1997). Automatic construction of news hypertext. In *Proceedings of Hypertext-Information Retrieval- Multimedia HIM97 Conference*, Dortmund, Germany.

[Furuta et al., 1989] Furuta, R., Plaisant, C., and Schneiderman, B. (1989). Automatically transforming regularly structured linear documents into hypertext. *Electronic Publishing*, 4(2):211–229.

[Golovchinsky and Chignell, 1997] Golovchinsky, G. and Chignell, M. (1997). The newspaper as an information exploration metaphor. *Information Processing and Management*, 33(5):663–683.

[Gower and Ross, 1969] Gower, J. C. and Ross, G. J. S. (1969). Minimum spanning trees and single linkage cluster analysis. *Applied Statistics*, 18:54–64.

[Hearst, 1994] Hearst, M. (1994). Multi-paragraph segmentation of expository texts. Technical Report TR94-790, Department of Computer Science, University of California Berkeley.

[Hearst and Plaunt, 1993] Hearst, M. and Plaunt, C. (1993). Subtopic structuring for full-length document access. In Khorfage, R., Rasmussen, E., and Willet, P., editors, *Proceedings of the 16th ACM-SIGIR conference*, pages 59–68, Pittsburgh, USA.

[Milligan and Cooper, 1985] Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50:159–179.

[Rada, 1992] Rada, R. (1992). Converting a book into hypertext. *ACM Transactions on Information Systems*, 10(3):294–315.

[Rumbaugh et al., 1991] Rumbaugh, J., Blaha, M., Premerlani, W., Eddy, F., and Lorensen, W. (1991). *Object Oriented Modeling and Design*. Prentice Hall.

[Salton et al., 1993] Salton, G., Allan, J., and Buckley, C. (1993). Approaches to passage retrieval in full text information systems. In Khorfage, R., Rasmussen, E., and Willet, P., editors, *Proceedings of the 16th ACM-SIGIR conference*, pages 49–58, Pittsburgh, USA.

[Salton et al., 1994a] Salton, G., Allan, J., and Buckley, C. (1994a). Automatic structuring and retrieval of large text files. *Communications of ACM*, 37(2):97–100.

[Salton et al., 1994b] Salton, G., Allan, J., Buckley, C., and Singhal, A. (1994b). Automatic analysis, theme generation, and summarization of machine-readable texts. *SCIENCE: Science*, 264.

[Salton and Buckley, 1989] Salton, G. and Buckley, C. (1989). Automatic generation of content links for hypertext. Technical Report TR89-993, Department of Computer Science, Cornell University, Ithaca, New York.

[Salton and McGill, 1983] Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

[Salton et al., 1995] Salton, G., Singhal, A., Buckley, C., and Mitra, M. (1995). Automatic text decomposition using text segments and text themes. Technical Report TR95-1555, Department of Computer Science, Cornell University, Ithaca, New York.

[Smeaton and Morrissey, 1995] Smeaton, A. and Morrissey, P. (1995). Experiments on the automatic construction of hypertext from texts. *The New Review of Hypermedia and Multimedia: Applications and Research*, 1.

[van Rijsbergen, 1979] van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworths, London.