Automatic Construction of News Hypertext

Theodore Dalamagas, Mark D. Dunlop
Computing Science Department
Glasgow University
Glasgow G12 8QQ, Scotland
E-mail: {dalamagt, mark}@dcs.gla.ac.uk

10 April 1997

Abstract

The automatic construction of hypertext is an important part of the hypertext authoring process. In this paper, we present a methodology for the automatic creation of 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 and Smeaton, 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* is not a trivial process.

In (Furuta et al., 1989), one of the earliest work 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) suggest 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 suggests, a methodology for the automatic construction of hypertext which is tailored to the domain of newspaper archives. In the second section, formal aspects of news hypertext are presented and modelling issues are discussed. In the third section we suggest methods to be used in order to automatically construct hypertext for a set of retrieved articles relevant to a query inserted by the user. Evaluation issues follow in the fourth section and, finally, directions of further work are discussed.

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. Remark 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 the 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 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 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

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

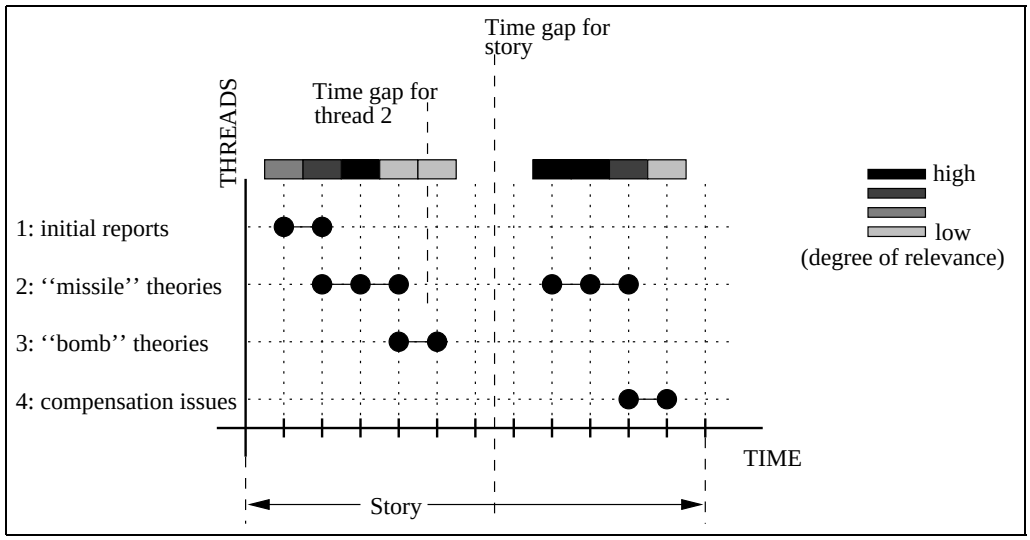

Figure 1: Temporal layout of story evolution

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* and T-links which connect them. We use the notion of segments 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[3]. 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 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.

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

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

[3] the algorithms to implement this will be discussed later

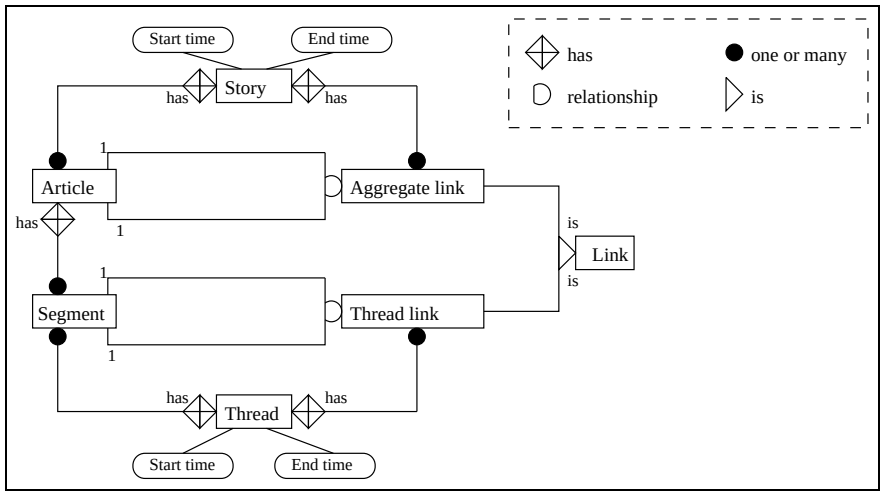

Figure 2: Object model

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 which are presented in the following sections, in contrast to the manual or thesaurus-based construction of concepts' set (Agosti et al., 1996).

3 A Methodology for the Automatic Construction of News Hypertext

3.1 Introduction

The automatic construction of A-links and T-links for news articles 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, our suggested methodology 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-Information Retrieval (NHIR)* system of figure 3.

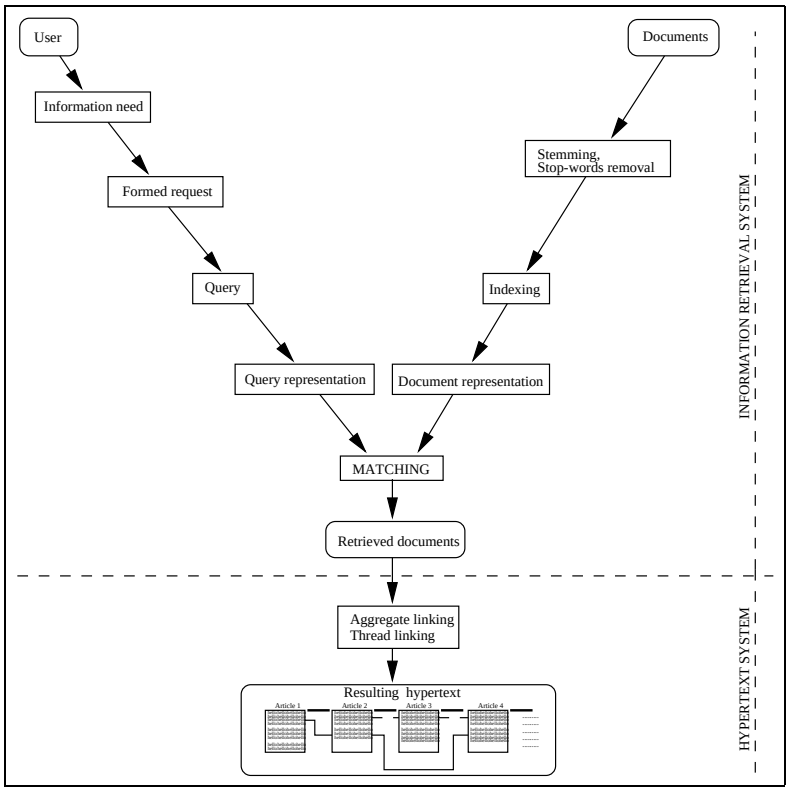

Figure 3: Architecture of NHIR system

Our 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.

However, as will be discussed later, certain aspects of the NHIR process can be carried out at index-time.

The components of the NHIR system are discussed in more details in the following sections: first the basic IR system and then the NHIR extensions.

3.2 Information Retrieval System

For the information retrieval system, each document (article) is represented by a set of terms (index terms) which are extracted from the full text of the document by using well-known techniques of stemming and stop-words removal (Frakes and Baeza-Yates, 1992). In the following paragraphs, the vector-space model (Salton and McGill, 1983) is briefly presented.

Following the vector-space model, p-dimensional document vectors $\mathbf{D}_i$ are constructed for each document i from a set of p index terms $t_1, t_2, ..., t_p$:

$$\mathbf{D}_i = (d_{i1}, d_{i2}, ..., d_{ip}) \tag{1}$$

where d_{ij} is the *weight* that is assigned to term j for document i. Term weights provide the degree of importance among terms for content representation of the document. A well known measure of term importance is obtained by using the product of term frequency and the *inverted document frequency (IDF)* (Salton and Buckley, 1988):

$$d_{ij} = tf_{ij} \cdot \log\left(\frac{N}{f_j}\right) \tag{2}$$

where

1. tf_{ij} is the frequency of term j in document i,
2. f_j is the number of documents that contain term t_j,
3. N is the number of documents,
4. j =1, 2, ..., p and
5. $\log\left(\frac{N}{f_j}\right)$ is the IDF of term j.

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

$$\mathbf{Q} = (q_1, q_2, ..., q_p) \tag{3}$$

where q_j is the *weight* that is assigned to term j for query Q. Salton et al. suggest the following query weighting formula:

$$q_j = \left(0.5 + \frac{0.5 \cdot tfq_j}{max(tfq)}\right) \cdot \log\left(\frac{N}{f_j}\right) \tag{4}$$

where

1. tfq_j is the frequency of term j in the query,

2. $max(tfq)$ is the maximum frequency of any term in the query.

Using the vector representations (1) and (3), similarity values are computed for each document-query pair:

$$Sim(\mathbf{D}_i, \mathbf{Q}) = \frac{\sum_{j=1}^{p}(d_{ij} \cdot q_j)}{\sqrt{\sum_{j=1}^{p} d_{ij}^2 \cdot \sum_{j=1}^{p} q_j^2}} \tag{5}$$

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 the information retrieval system is a set of articles (documents) related to the topic that is determined by the user with the form of a query.

3.3 Hypertext Construction System

Creating A-links between the articles 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 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. (Salton et al., 1994b; Salton et al., 1995) and is described as follows:

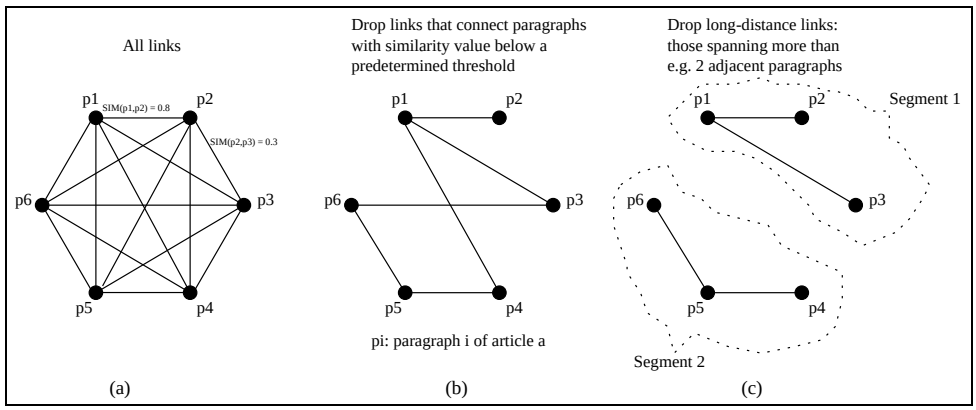

Figure 4: Simplification of text relationship map for segment detection

1. Construct the *paragraph relationship map* of the article that needs to be decomposed into segments. This map has the form of a graph. Its vertices

correspond to the paragraphs of the article 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 0.2 between paragraphs $p2$ and $p3$ shows that these paragraphs are not related.

2. Drop the links of the map that correspond to similarity values which are below a predetermined threshold (see figure 4b).

3. Drop long-distance links: those spanning more than a predetermined number of adjacent paragraphs (see figure 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. Using this process, all the articles of a collection can be decomposed into segments at index-time, prior to any usage of the NHIR system. After this process, n_i segments exist for each article a_i:

$$a_i \longrightarrow (S_{i1}, S_{i2}, ..., S_{in_i}) \tag{6}$$

Having the set of all segments S of all articles, segment clustering can be performed for the set of retrieved articles relevant to query that create a story. Each formed cluster is used in order 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 5.

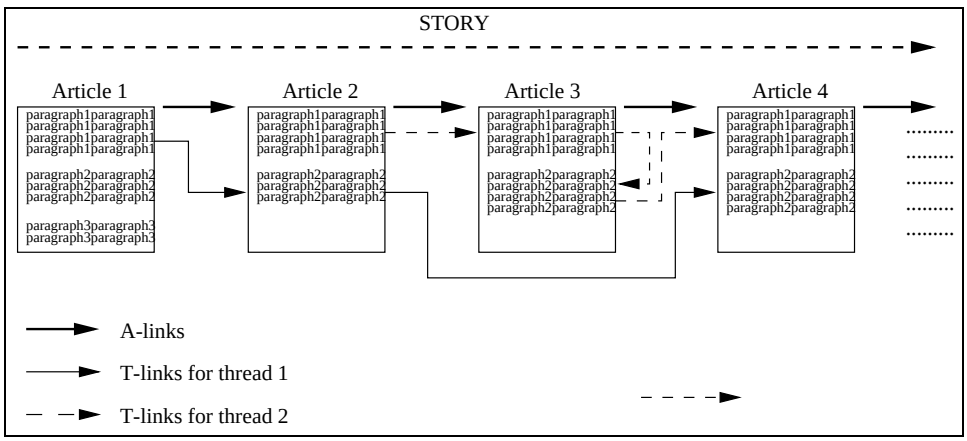

Figure 5: News hypertext

Despite the fact that the articles in figure 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 the 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 extract segments that may be related to each other, although they are separated in the article. Clustering will also group together these segments.

From a wide variety of clustering algorithms (see (Willet, 1988) for a great overview), the single-link method (Sibson, 1973; van Rijsbergen, 1971) is an extensively used hierarchical method in information retrieval (van Rijsbergen, 1979). However, the single-link method, like all hierarchical methods, has time requirements of $O(n^2)$ (if n documents need to be clustered), because all pairs of similarities must be considered[4]. As a result, the performance overhead limits its usefulness for on-the-fly clustering. Non-hierarchical clustering methods have lower computational requirement, $O(kn)$ (if n documents need to be grouped into k clusters), but certain parameters like the number of formed clusters must be known a priori.

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 (Willet, 1988) 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.

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. Experiments will be conducted in order to select the most effective clustering method.

Allan (Allan, 1995) suggests to use the similarity matrix between document segments and put links between segments that are related with high similarity value. This approach does not take into consideration the temporal aspects of news hypertext and the notion of story evolution. We believe that the clustering approach which we suggest 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.

4 Evaluation

Information retrieval evaluation techniques, like precision and recall graphs, cannot be directly used in the evaluation of an NHIR system (Agosti and Smeaton, 1996). We suggest that the presented methodology for the automatic construction of news hypertext should be evaluated from three different aspects:

[4]Willet (Willet, 1980) suggests an inverted file algorithm to limit the amount of computation required to calculate a similarity matrix

- *Hypertext structure analysis*
  Various metrics have been developed for structural analysis of hypertext. *Compactness*, the intrinsic connectedness of the hypertext, and *stratum*, the degree to which the hypertext is organized so that some nodes must be read before others, can be used to measure the overall topology of the 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 (Agosti and Smeaton, 1996). 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 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 clustering results determine the threads of articles that refer to substories within a 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 occuring 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).

- *System usability*
  The effectiveness of the system that will implement our proposed framework needs also to be tested from the user's point of view. We believe that the time that a user spends in order to fulfil her information needs is a critical measure for the effectiveness of the NHIR system. Therefore, a comparison between a classical IR system and the NHIR 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 involves 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 the evaluation of hypertext information retrieval one can refer to (Dunlop, 1996).

SMART retrieval system (Salton and McGill, 1983; Buckley, 1985) has been selected as the platform in which the suggested framework will be implemented and

tested[5]. Although SMART is an academic research software and it is not optimized for any particular usage, it is designed with great flexibility. The document collection that will be used consists of 23235 articles (140MB) 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.

5 Further Work

As this work is currently in progress, there are many issues that need to be considered in detail. Among them, the most important are discussed in the following paragraphs:

- *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 will form small threads within the main story chain and thus they can be eliminated. The idea has been applied successfully in (Hearst and Pedersen, 1996).

- *Handling time gaps*
  In the 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*
  In figure 1, the evolution of story is presented in a temporal layout together with its threads. Thread surrogates are necessary for the understanding of the story evolution. A surrogate of thread may be a title, a list of keywords, a summary, etc. As a result, techniques involving keyword extraction, theme identification and summarizing should also be considered.

- *Estimating parameters*
  Using similarity values in order to determine content relations between texts requires the right setting of the threshold parameters. Dealing with articles that have great probability to be related to each other, because they are considered to be relevant to the same query, makes the estimation of threshold parameters difficult. Parameters need to be set for the decomposition process, too. Segment formation is based on the drop of links, spanning more than α adjacent paragraphs, where α is pre-determined. Using fixed or adjustable

[5] *Although experimentation is still underway, and hence cannot be reported here, it is our intention to include some results in the final version of this paper, if accepted, and present all the results at the conference in October.*

parameters is another important issue. The former have the advantage of simplicity but they might lead to vague results, especially in the case of text decomposition. In contrast, making for example the α parameter proportional to the length of article, segment formation could be more accurate.

- *Performance*
  The creation of hypertext after the process of article retrieval imposes an overhead on system's performance. However, on-the-fly clustering has been also used in (Hearst and Pedersen, 1996) without great decrease of performance because of the small number of texts that deals with (comparing to the whole article collection) and the fast clustering algorithm that has been used (see section 3.3). The performance will be monitored during the users' test.

6 Conclusion

In this paper, we presented a methodology for the automatic construction of news hypertext. News hypertext has an important characteristic comparing with textbook hypertext: time should be also taken into account as an attribute of the hypertext. Thus, we consider the presence of chronologically ordered chains of linked texts, related to a topic, to be of major value.

In order to satisfy this special requirement of news hypertext, we suggest the notion of a story. An NHIR 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. The hypertext construction is done on-the-fly and only for the set of the retrieved articles.

Classic hierarchical clustering cannot be used directly for on-the-fly clustering due to its performance overhead. However, a combination of hierarchical and non-hierarchical methods has given promising results. In any case, the choice of an appropriate fast clustering algorithm is of great importance for the quality of the created T-links and the system performance. Experiments will be conducted to select the most effective clustering method.

Finally, the presented methodology for the automatic construction of news hypertext should be evaluated in terms of the hypertext structure, hypertext quality and system usability. The first two criteria have strong mathematical foundations and are based on various metrics. The third one is based on human-computer interaction techniques in which a user performs tasks to fulfil her information needs.

Acknowledgement

This paper is part of the work for Theodore Dalamagas' MSc thesis in Advanced Information Systems course in the Computing Science Department of Glasgow University. The work is supervised by Mark D. Dunlop.

References

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.

Agosti, M. and Smeaton, A., editors (1996). *Information Retrieval and Hypertext.* Kluwer Academic Publishers.

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

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

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

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.

Dunlop, M. D., editor (1996). *Proceedings of the Second Mira Workshop*, Monselice, Italy. University of Glasgow Computing Science Research Report TR-1997-2, http://www.dcs.gla.ac.uk/mira/workshops/padua_procs.

Frakes, W. and Baeza-Yates, R., editors (1992). *Information Retrieval: Data Structures and Algorithms.* Prentice Hall.

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

Hearst, M. and Pedersen, J. O. (1996). Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 76–84, Zurich, Switzerland.

Ling, R. F. and Killough, G. G. (1976). Probability tables for cluster analysis based on a theory of random graphs. *Journal of the American Statistical Association*, 71:293–300.

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

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

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

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

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.

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

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.

Sibson, R. (1973). SLINK: An optimally efficient algorithm for the single-link cluster method. *Computer Journal*, 16:30–34.

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, C. J. (1971). An algorithm for information structuring and retrieval. *Computer Journal*, 14:407–412.

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

Willet, P. (1980). Document clustering using an inverted file approach. *J. Information Science*, 2:223–231.

Willet, P. (1988). Recent trends in hierarchic document clustering: A critical review. *Information Processing and Management*, 24(5):577–597.