To be presented at the ICCC/IFIP 5th Conf. on Electronic Publishing ELPUB'01, Canterbury,UK, 5-7 July 2001

A framework for automatic combination of media contents by minimising information redundancy Case: Integrated publishing in multimedia networks

Anneli Heimbürger (*), Paula Silvonen and Caj Södergård
VTT Information Technology
P.O. Box 1204
FIN - 02044 VTT
Finland
(*)*Email: anneli.heimburger@vtt.fi*

Abstract

Information redundancy becomes a crucial problem in the Web when contents from different resources are automatically combined to produce a new WWW–publication. Information retrieval, natural language processing and the latest WWW–activities offer a challenging framework to approach the information redundancy problem of automatically combined news articles. It seems reasonable, that minimising information redundancy should be performed by a hybrid technique that combines some elements of these approaches. The purpose of this exploratory study is to introduce a theoretical and practical framework for clarifying the information redundancy problem in the case of integrated publishing.

1. Introduction

For millions of Internet users, resource discovery from the World Wide Web would be a more fascinating and efficient experience if there were automatic or at least semiautomatic methods of detecting and filtering out information redundancy. The information redundancy problem becomes crucial in dynamic, distributed Web database applications in which heterogeneous information is automatically integrated from different data resources. One such application area is integrated publishing, where materials coming from several Web newspapers and TV news are automatically combined to form a new integrated Web–publication or a news service.

If the Web-based news service provides a user with a set of related documents that might overlap to a certain extent, the user is probably interested in the union of these news articles where similarities are eliminated. In this paper, news is defined as information about recent national and international events of general interest currently reported by

news bureaux, newspapers, radio, television and Web news services. The focus of this exploratory study is on textual news items.

The exploratory study presented here is a part of the integrated multiple media publishing research and development work currently in progress in VTT Information Technology, one of eight research institutes that make up the Technical Research Centre of Finland (VTT) (Södergård et al. 1999). The case environment of the study is the Web-based integrated publication system for TV–programmes and newspaper articles called IMU (http://www.vtt.fi/imu2). IMU is being developed by VTT Information Technology together with Finnish universities and enterprises. The IMU system has been designed to work with mobile devices as well. Minimising the information redundancy of news articles is of special importance in mobile news services.

The IMU system has two channel types: common and personal channels (Figure 1). The channel is defined as a news composite. The common channel offers general news items and the personal channel can be personalised according to user's interests. From the point of view of an IMU user, one ideal solution would be to provide a user with a short summary concerning a certain news topic. The summary could be generated or extracted automatically from the IMU system, and it could also have a list of links to articles in different newspapers and TV news stories dealing with the topic concerned. We could define this kind of a summary as a meta-summary or a meta-ingress. However, in the current IMU-implementation the common and personal news channels overlap to some extent. The same or almost the same news story can appear several times with different or slightly different titles depending on how the newspapers have dealt with the topic and how sensational the news topic has been. The main reason for this is that the newspapers base their stories on common news items from national and international news bureaux.

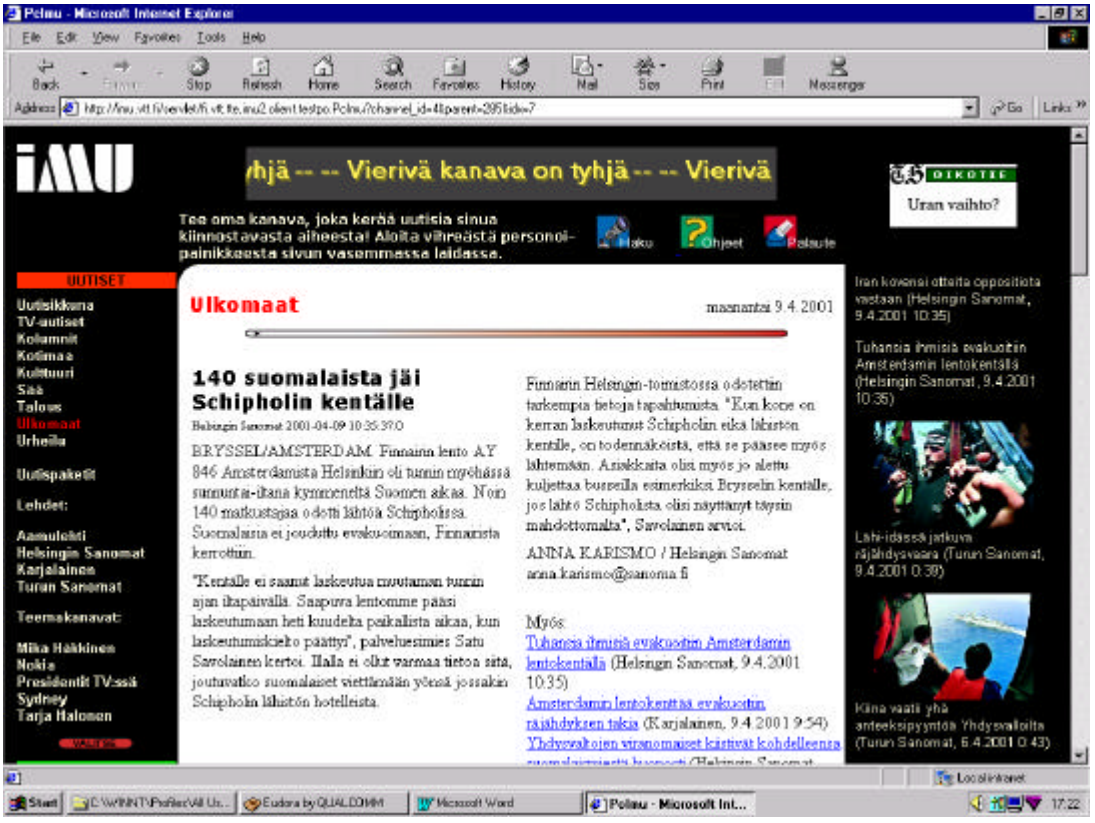

Figure 1. The IMU system has two channel types on the left: common and personal channels. The common channel offers general news items and the personal channel can be defined by the user according to his/her interests.

Minimising information redundancy can be divided into two phases: detecting redundancy and filtering out redundancy. A step-by-step approach to solve the problem seems a reasonable way to start. The process of determining whether one news article is similar to another one involves two main steps: (a) extracting appropriate information from the incoming news article and (b) comparing its features against those previously extracted from news articles in the database. What features to use and how they are compared are the two primary issues to be resolved. The more reliable the feature extraction is, the better the result. In terms of computational efficiency the feature-comparing –step must be very fast if the database is large and real-time performance is desired. After semantic similarity is somehow and on some level detected, the articles should be clustered around different topics, and meta-ingresses generated or extracted inside the topic clusters.

There are several scenarios when the elimination of information redundancy can be useful. For example, Web search engines often retrieve identical or nearly identical documents. A redundancy detection mechanism could be used as a filter for search results, excluding documents that overlap with documents above a certain threshold. Another application could be to highlight the differences between these documents. The goal of automatic information extraction and integration techniques is to construct combined descriptions of the information coming from multiple heterogeneous resources.

Classic and modern information retrieval, natural language processing, the latest WWW–activities, journalistic and communication research, and usability engineering provide a multidisciplinary approach to solving the problem of information redundancy in news articles. In this study we concentrate on the first three issues. It seems obvious that detecting and filtering out information redundancy should be performed by a hybrid technique that combines elements of these issues.

There are several research projects working on electronic news delivery systems (see Section 4). As far as we know our approach differs from others in two ways:

- In the IMU system the integration of news articles from several Web resources is automatic.
- The detection and filtering out of information redundancy focuses on the semantic metadata of news articles, which can be automatically extracted from the relational database of the IMU system.

The purpose of this exploratory study is:

- to introduce elements of the framework for clarifying and formalising the information redundancy problem in the case of integrated publishing,
- to describe experimental arrangements to test the detection and filtering out of information redundancy,
- to identify relevant problems for more precise investigation.

The remainder of this paper is organised as follows. A framework for approaching the information redundancy problem of news articles is introduced in Section 2. The working hypothesis for the experimental arrangements is described in Section 3. Related work is reviewed in Section 4. Future work is discussed in Section 5.

2. Elements for the theoretical framework

Information retrieval together with natural language processing offer some methods to approach the information redundancy problem. Recent developments in World Wide Web techniques, such as the Resource Description Framework (RDF), the Topic Maps –standard, ontologies and the concept of the Semantic Web as a whole are essential tools to

be considered for automatic metadata extraction and semantic markup generation. All of these provide interesting possibilities for developing the Web news services of the future.

2.1 Information retrieval

The main objective of information retrieval research is to develop methods for retrieving all the documents which are relevant to a user query. "Find related/similar" and duplicate detection –types of algorithms are interesting from the viewpoint of our research problem.

Clustering methods with distance measures can be used to reduce information by grouping together similar items, and to generate simplified descriptions and summaries of documents. An optimal clustering method loses as little relevant information as possible discussed in documents. Combining different distance measures may provide a new approach to similarity detection. This is one issue for our future work.

2.2 Natural Language Processing

Natural language gives freedom for enormous variation in expression, from choosing between synonyms to using different styles, emphasis, different levels of abstraction, and metaphoric expressions. Authors have their own individual writing styles, which depend on their background, knowledge, profession and personal style of communication. One basic problem is that the same idea can be presented in several ways.

In a useful text analysis method, synonymous expressions should be encoded similarly. The goal of natural language processing is to use the semantic information of documents as well as statistical information to enhance the clustering of documents. In our research we use a string-matching algorithm with a thesaurus (see Section 4).

2.3 Semantic Web

The Semantic Web is an idea of World Wide Web inventor Tim Berners-Lee, Director of the World Wide Web Consortium (Berners-Lee 1999; W3C 2001). The word "semantic" in the context of the Semantic Web means "machine-processable". Berners-Lee explicitly rules out the sense of natural language semantics. The aim of the Semantic Web is to have data on the Web defined and linked in a way that it can be used by machines not just for display purposes, but for integration and reuse of information across various applications.

Markup languages define the markup rules, which add meaning to the structure and content of documents. The Standard Generalised Markup Language (SGML) is an international standard for defining markup languages, and it is designed to promote text interchange (ISO 8879 1986). The Extensible Markup Language (XML) is a subset of the SGML, and it is usable over the Internet. Like SGML, an XML document consists of marked up data using tags. XML documents can have a document type definition, but it is not necessary. The textual contents of IMU news articles are stored in the XMLNews-Story format that defines the structural components of a news article (XMLNews-Story Specifications 2001).

The Resource Description Framework (RDF) is another application of XML (W3C 2001). RDF is a general framework for describing any Internet resource. Such descriptions are often referred to as metadata or "data about data". An RDF description can include the authors of the resource, the date of creation or updating, key words, subject categories, and so forth. The RDF language is an important schema for the description of resources and their types. Recent developments in markup allow domain-specific information to be expressed as document metadata.

Above this, according to the Berners-Lee Semantic Web architecture, there is the ontology layer. An ontology is a set of concepts - such as things, events and relations - that are defined, for example, in domain-specific controlled language in order to create an

agreed-upon vocabulary for exchanging information (Guarino 1995). To standardise semantic terms, many areas use specific ontologies, which are hierarchical taxonomies of

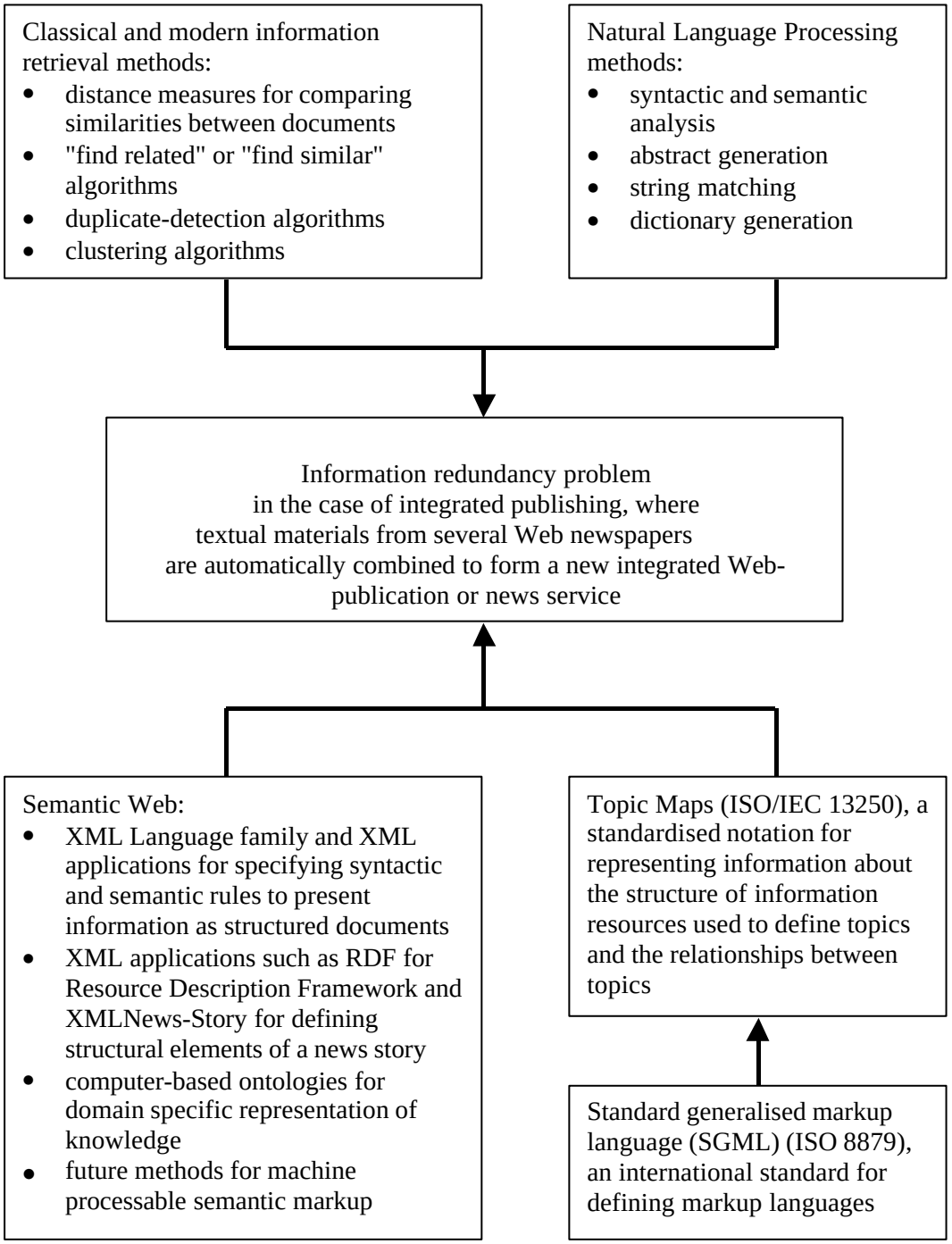

Figure 2. The framework for approaching the information redundancy problem in the case of integrated publishing.

terms describing certain knowledge topics (Crow and Shadbolt 1996; Miller 1995; Motta et al. 2000; Staab et al. 2000). In the IMU system the classification of news into basic topics can be seen as a very rough ontology.

The Topic Map –standard might provide an interesting possibility to define a machine-processable ontology for the IMU news service. Topic Maps are defined by the ISO/IEC 13250 standard (ISO/IEC 13250 2000). The base notation of Topic Maps is

SGML. As XML is a subset of SGML, it can be also used as a base notation for Topic Maps. A topic map expresses its creator's view about what the topics are, and which parts of an information set are relevant to which topics.

Finally, we put these theories, models and methods in perspective and collect them into a unified framework. The introduced framework provides us with methods to approach the information redundancy problem and to construct the experimental environment for testing. It also gives us ideas for future work as well (Figure 2).

3. Towards a practical framework

The case environment of our experimental research on minimising information redundancy will be based on the integrated publication system for TV–programmes and newspaper articles called IMU (http://www.vtt.fi/imu2). A continuously updated WWW-multimedia publication is created by combining the contents of databases from several newspaper houses and TV companies. The users read the news service from the Internet on their PC's with normal WWW-browsers, or from television sets connected to the Internet through set-top –boxes. The content can be accessed with WAP phones. The news can also be downloaded as synthesised speech into portable MP–players. This is a service of special interest for the visually impaired.

The heart of the IMU system is an active media server. Articles from four Finnish newspapers and TV news from one Finnish company are deposited in its database. The IMU system analyses the Web version of the newspaper, divides it up into structural elements, and separates off the necessary metadata. In the same way, the TV news is broken down into news items, using video analysis and closed caption texts. The contents are deposited in the database as objects, which are then combined to form a publication tailored to the reader's preferences.

At the moment the IMU consists of three types of information as follows, each type with related descriptive and/or semantic metadata.

- television news broadcasts: video and metadata. A video news article containing text, picture and metadata is generated for each news item within the news broadcast.
- news articles: text, pictures and metadata
- events and programme information: text, metadata like locations, start and end times.

The textual content of articles, images and video files are stored in the file system. The textual content of articles is stored in the XML-based XMLNews-Story format. The metadata of an article is stored in the relational database.

We divide a Web news article into separate components at different levels of precision. A textual news article N on the Web can be presented as follows:
Level 0: N = (n),
Level 1: **N** = (M, C),
Level 2: **N** = (**M**, C),
Level 3: **N** = (**M**, **C**),
Level 4: **N** = (**M**, A, **C**),
Level 5: **N** = (**M**, A, **C**, **L**), where

- n is the news article as a whole
- M is for metadata and C is for content
- **M** = (m_1, m_2), m_1 for descriptive metadata and m_2 for semantic metadata. Descriptive metadata describes, for example, by whom, how and when documents were created. Semantic metadata characterises the actual subject matter that can be found within the document's contents.
- **C** = (c_1, c_2, ... c_n) is for content where c_i, i = 1, 2, ... n are structural elements of the news story, for example, according to the XMLNews-Story DTD

- A is for an abstract.
- **L** = (l_1, l_2) is for links, where l_1 represents outgoing links and l_2 incoming links.

We could continue the analysis further by taking account of visual information (graphics, stills, animation and video) and audio information (speech, music, voice and other sounds) materials with related time dependence and with different types of links etc. However, this is not done in this work. The analysis will help us to identify different components and combinations of components on which the information redundancy detection and filtering can be focused, and to generate ideas for experimental arrangements.

Demonstrations of minimising information redundancy in the automatic combination of media contents will focus in the IMU case on textual information in the first place. In the Figure 3 our working hypothesis for experimental tests is presented. The semantic metadata elements such as titles, photo captions, article descriptors and ingresses are automatically extracted from incoming news articles, which are stored on the IMU active server and tagged according to the XMLNews-Story specification. Semantic similarity analysis will be carried out in the first phase between article titles, in the second phase between article photo captions, in the third phase between article descriptors, and finally, if necessary, between article ingresses. The predefined value of a similarity threshold function will define whether the process of the similarity analysis proceeds to the next phase.

The similarity calculations will be performed using the UNIFIER algorithm, which was developed at VTT Information Technology. Unifier is a Java-based software that analyses string similarity by applying the minimum edit-distance metric and dictionary lookup. The system can be utilised, for instance, in analysing the content similarity of text documents and in spelling error correction. The minimum edit-distance metric is expanded to whole words instead of only characters (Jurafsky and Martin 2000). In addition to minimum edit-distance calculation, a thesaurus is utilised in phrase similarity analysis. Thesaurus entries define synonyms that do not look alike and so cannot be matched with the edit-distance method.

The edit-distance or Levenshtein distance is defined in our approach using the following types of string difference:

- Substitution – one letter or word is replaced by a different one (*peoole* vs. *people*; on phrase level *Bond 007* vs. *Bond James*)
- Deletion – one letter or word is missing from the string (*peole* vs. *people*; *Bond* vs. *Bond James*)
- Insertion – one extra letter or word is inserted into the string (*peopple* vs. *people*; *Bond, James Bond* vs. *Bond James)*
- Transposition of two adjacent letters or words (*pepole* vs. *people*; *James Bond* vs. *Bond James*)

The system considers these string distance types and their combinations when it calculates similarity measures both for words and phrases.

After the similarity analysis, article elements with pointers to articles will be clustered around topics. If the news stream comes from the personal channels, then the topics are already defined by a user. In the case of the common news channel the topic clusters must be generated. The news topics can be the main categories of the IMU system or a more precise classification in which ontological methods and the Topic Maps –standard could be applied. Links to other articles inside the cluster should also be created.

Then we have to decide inside one topic cluster which news ingress we show for a user. The decision can be made according to the Euclidean distance in the cluster. There are also other approaches, such as a time stamp priority and a front-page story priority.

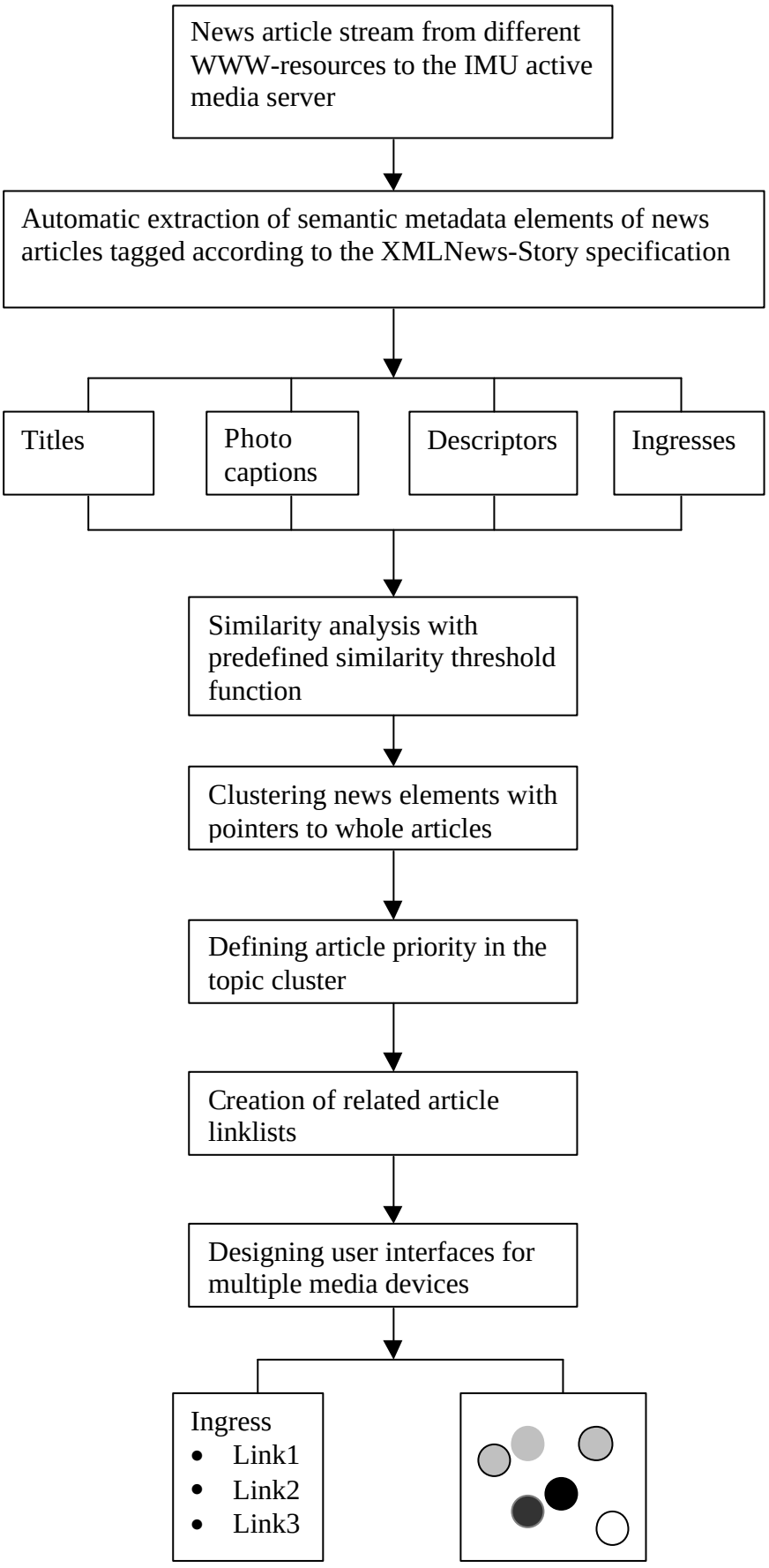

Figure 3. A practical framework to test the detection and filtering out of information redundancy.

Finally, the user interfaces for multiple media devices should be designed. For mobile phones the textual interface is a natural choice. For PCs and other mobile devices such as communicators, there are more degrees of freedom to visualise news clusters and the similarity relations between news articles by graphical views.

4. Related work

There are several research projects working on electronic news delivery systems. To the best of our knowledge our approach differs from others in that we focus on the information redundancy problem and our approach is based on the semantic metadata of XMLNews-Story tagged news articles, which can be automatically extracted from the relational database of the IMU system. Most of the research on electronic news delivery systems concentrates on searching news (Hunter 2001; Maybury 2000; Mock and Vemuri 1997; Watters and Wang 1998, 2000).

Information retrieval research provides a set of concepts, models and methods that seem to be useful when approaching the information redundancy problem in our case. One very good starting point for information retrieval research is the home page of the Center for Intelligent Information Retrieval (CIIR) (http://ciir.cs.umass.edu/). There are a number of issues such as information extraction, interfaces and visualisation, topic detection and tracking being studied with the general aim of improving the effectiveness of access to and organisation of large text databases (Croft 2000). CIIR's topic detection and tracking (TDT) research project has investigated the state of the art in finding and following new events in a stream of broadcast news stories. The goal of the research is to organise television, radio and print news according to the event-based topics that they discuss.

Information extraction (IE) systems analyse unrestricted text in order to extract specific types of information. IE research has been reported by, among others, Appelt (1999), Cowie and Lehnert (1996), Endres-Niggemeyer (2000), Fuketa et al. (2000), Liu et al. (1999), Moens and Dumortier (2000), and Salton et al. (1997). Automatic text categorisation and summarisation has potential in many text-based applications including text routing and filtering (Lam et al. 1999; Moens and Dumortier 2000; Salton et al. 1997). From a natural language perspective, information extraction systems must operate at many levels, starting from word recognition and sentence analysis and ending up at comprehension of the full text document (Jurafsky and Martin 2000).

Much research has been done on similarity measures and clustering (Baeza-Yates and Ribeiro-Neto 1999; Kohonen 2001; Kok et al. 1999; Kowalski and Maybury 2000; Modha and Spangler 2000; Skarmeta et al. 2000; Zhang and Korfhage 1999). Carrick and Watters (1997) and Watters et al. (1998) have studied the automatic association of text and photo news items in their Electronic News Delivery -project. Bergamaschi et al. (2001) have studied semantic similarity and calculation of the semantic similarity degrees in the context of integrating data from heterogeneous database systems.

Many systems have been proposed to find similar or related documents in the World Wide Web (Dean and Henzinger 1999; Hui and Goh 1998; Netscape Communications Corporations 2001; Vu and Li 2000).

Intellectual property protection and work on duplicate detection has concentrated mostly on exploring extracted features of document image databases (Brin et al. 1995; Lopresti 2000) or CAD applications (Kahng et al. 1999). Monostori et al. (2000) have studied methods for detecting plagiarism in large text collections.

There are also interesting projects going on in the EU's Information Society Technologies (IST) –programme such as the Multimedia Indexing and Searching Environment (MUMIS). The project focuses on technology for indexing and retrieval of data from different media sources and in different languages (Hiddink 2001).

5. Discussion and future work

The exploratory study has concentrated on the methodological and technical issues related to minimising information redundancy. In this study the theoretical framework for minimising the information redundancy problem in the case of automatic combination of media contents in integrated publishing has been introduced. The hybrid framework contains elements from information retrieval research, natural language processing and the latest developments of the World Wide Web. The experimental arrangements in the IMU environment have been described as a practical framework of the study.

In this study the following problems were identified for further research:

- a better understanding of the possibilities of existing metadata in the IMU relational database,
- a more precise understanding and identification of different information redundancy categories in the common and personal channels of the IMU system,
- a systematic analysis of the effect of the similarity threshold function value to the results,
- a better understanding of the characteristics of news materials and the effects of their time dependence on the study of information redundancy,
- a more precise understanding of distance measures and their possible combinations that could have the potential to provide a new approach to similarity research,
- a more precise understanding of the role of ontologies, the Topic Map –standard and semantic markup methods in the conceptualisation and management of media contents,
- a systematic analysis of the information redundancy problem in mobile news services.

Furthermore, the increased availability of multi-lingual materials in all areas of the public and private sector is driving demand for systems that facilitate integration of the contents from multi-lingual Web resources. Developing methods for minimising news material redundancy in automatic combination of multi-lingual media contents in mobile news services challenges the future of the Semantic Web.

References

Appelt, D. E. 1999. Introduction to information extraction. AI Communications, Vol. 12, No. 3, pp. 161-172.

Baeza-Yates, R. and Ribeiro-Neto, B. 1999. Modern information retrieval. New York: ACM Press. 513 p. ISBN 0-201-39829-X.

Bergamaschi, S., Castano, S., Vincini, M. and Beneventano, D. 2001. Semantic integration of heterogeneous information sources. Data & Knowledge Engineering, Vol. 36, No. 3, pp. 215 - 249.

Berners-Lee, T. 1999. Weaving the Web. New York: HarperSanFrancisco. 226 p. ISBN 0-06-251586-1.

Brin, S., Davis, J. and García-Molina, H. 1995. Copy detection mechanisms for digital documents. In: Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, May 22 - 25, 1995, San Jose, CA, USA. Pp. 398 - 409.

Carrick, C. and Watters, C. 1997. Automatic association of news items. Information Processing and Management, Vol. 33, No.5, pp. 615 - 632.

Cowie, J. and Lehnert, W. 1996. Information extraction. Communications of the ACM, Vol. 39, No. 1, pp. 80 – 91.

Croft, W. B. 2000. Recent research from the Center for Intelligent Information Retrieval. Boston: Kluwer Academic Publishers. 306 p. ISBN 0-7923-7812-1.

Crow, L. and Shadbolt, N. 2001. Extracting focused knowledge from the semantic web. International Journal of Human-Computer Studies, Vol. 54, No. 1, pp. 155 - 184.

Dean, J. and Henzinger, M. R. 1999. Finding related pages in the World Wide Web. Computer Networks, Vol. 31, No. 11 - 16, pp. 1467 - 1479.

Endres-Niggemeyer, B. 2000. SimSum: an empirically founded simulation of summarizing. Information Processing and Management, Vol. 36, No. 4, pp. 659 - 682.

Fuketa, M., Lee, S., Tsuji, T., Okada, M. and Aoe, J-I. 2000. A document classification method by using field association words. Information Sciences, Vol. 126, No. 1-4, pp. 57 - 70.

Guarino, N. 1995. Formal ontology, conceptual analysis and knowledge representation. International Journal of Human and Computer Studies, Vol. 43, No. 5/6, pp. 625 - 640.

Hiddink, W. 2001. MUMIS Multimedia Indexing and Searching Environment. (referred 4.4.2001). Project description: <URL: http://www.ctit.utwente.nl/Docs/projects/international/mumis.html>.

Hui, S-C. and Goh, A. 1998. Information filtering of on-line news using dynamic abstract generation. Cybernetics and Systems, Vol. 29, No. 6, pp. 577 - 591.

Hunter, A. 2001. Get your news on the Web. Information World Review, January 2001, No. 165, pp. 20 - 21.

ISO 8879. 1986. Information Processing - Text and office systems - Standard generalised markup language (SGML). Geneva: ISO. 155 p.

ISO/IEC 13250. 2000. Information Technology - SGML Applications - Topic Maps. Geneva: ISO/IEC. 39 pp.

Jurafsky, D. and Martin, J. H. 2000. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River: Prentice-Hall, Inc. 934 p.

Kahng, A. B., Kirovski, D., Mantik, S., Potkonjak, M. and Wong, J. L. 1999. Copy detection for intellectual property protection of VLSI designs. In: Proceedings of the 1999 International Conference on Computer-aided Design. Pp. 600 - 605.

Kohonen, T. 2001. Self-organizing maps. Third edition. Berlin: Springer (Springer series in Information Sciences 30). 501 p. ISBN 3-540-67921-9.

Kok, Y. H., Goh, A. and Holaday, D. A. 1999. Using cluster analysis to determine the media agenda. Aslib Proceedings, Vol. 51, No. 10, pp. 361 - 371.

Kowalski, G. J. & Maybury, M. T. 2000. Information storage and retrieval systems. Theory and implementation. Second edition. Boston: Kluwer Academic Publishers. 318 p. ISBN 0-7923-7924-1.

Lam, W., Ruiz, M. and Srinivasan, P. 1999. Automatic text categorization and its application to text retrieval. IEEE Transactions on Knowledge and Data Engineering, Vol. 11, No. 6, pp. 865 - 879.

Liu, J., Wu, Y. and Zhou, L. 1999. A hybrid method for abstracting newspaper articles. Journal of the American Society for Information Science, Vol. 50, No. 13, pp. 1234 - 1245.

Lopresti , D. P. 2000. String techniques for detecting duplicates in document databases. International Journal on Document Analysis and Recognition, Vol. 2, No. 4, pp 186 - 199.

Maybury, M. 2000. News on demand. Communications of the ACM, Vol. 43, No. 2, pp. 33 - 34.

Miller, G. A. 1995. WordNet: A lexical database in English. Communications of the ACM, Vol. 38, No. 11, pp. 39 - 41.

Mock, K. J. and Vemuri, V. R. 1997. Information filtering via hill climbing, wordnet, and index patterns. Information Processing & Management, Vol. 33, No. 5, pp. 633 - 644.

Modha, D. S. and Spangler, W. S. 2000. Clustering hypertext with applications to web searching. In: Proceedings of the 11th ACM on Hypertext and Hypermedia, May 30 - June 3, 2000, San Antonio, TX USA. Pp. 143 - 152.

Moens, M-F. and Dumortier, J. 2000. Text categorization: the assignment of subject descriptors to magazine articles. Information Processing & Management, Vol. 36, No. 6, pp. 841 - 861.

Monostori, K., Zaslavsky, A and Schmidt, H. 2000. Document overlap detection system for distributed digital libraries. In: Proceedings of the 5th ACM conference on ACM 2000 Digital Libraries, June 2 - 7, 2000, San Antonio, TX, USA. Pp. 226 - 227.

Motta, E., Buckingham Shum, S. and Domingue, J. 2000. Ontology-driven document enrichment: principles, tools and applications. International Journal of Human-Computer Studies, Vol. 52, No. 6, pp. 1071 - 1109.

Netscape Communications 2001. What's Related FAQ (referred 16.3.2001) <URL: http://home. netscape.com/escapes/related/faq.htm>.

Salton, G., Singhal, A., Mitra, M. and Buckley, C. 1997. Automatic text structuring and summarization. Information Processing & Management, Vol. 33, No. 2, pp. 193 - 207.

Staab, S., Angele, J., Decker, S., Erdmann, M., Hotho, A., Maedche, A., Schnurr, H.-P., Studer, R. and Sure, Y. 2000. Semantic community Web portals. In: Proceedings of the 9th International World Wide Web Conference, Amsterdam, The Netherlands, Elsevier Amsterdam. Pp. 473 - 491.

Södergård, C., Aaltonen, M., Hagman, S., Hiirsalmi, M., Järvinen, T., Kaasinen, E., Kinnunen, T., Kolari, J., Kunnas, J. and Tammela, A. 1999. Integrated multimedia publishing: combining TV and newspaper content on personal channels. Computer Networks, Vol. 31, No. 11-16, pp. 1111 - 1128.

Vu, Q. and Li, W-S. 2000. Exploring link topology for associating Web pages. Poster in the 9th International World Wide Web Conference, Amsterdam, May 15 - 19, 2000. 3 pp.

W3C The Technology & Society Domain: Semantic Web Activity (referred 12.3.2001) <URL:http://www.w3.org/2001/sw/>.

Watters, C. and Wang, H. 2000. Rating news documents for similarity. Journal of the American Society for Information Science, Vol. 51, No. 9, pp. 793 - 804.

Watters, C., Shepherd, M. A. and Burkowski, F. J. 1998. Electronic news delivery project. Journal of the American Society for Information Science, Vol. 49, No. 2, pp. 134 - 150.

XMLNews-Story Specifications (referred 9.4.2001) <URL: http://www.xmlnews.org/XMLNews/>.

Zhang, J. and Korfhage, R. 1999. Distance and angle similarity measure method. Journal of the American Society for Information Science, Vol. 50, No. 9, pp. 772 - 778.